

A HIGH-EFFICIENCY VLSI CNN ACCELERATOR FOR REAL-TIME BIOANALYTICAL IMAGING ON EDGE DEVICES

G.Chandana
Dept. of ECE
C. Byregowda Institute of Technology (CBIT)
Kolar, Karnataka, India
Chandana7433@gmail.com,

Pallavi K.V
Dept. of ECE
C. Byregowda Institute of Technology (CBIT)
Kolar, Karnataka, India
Pallavikv339@gmail.com

Aishwarya.M
Dept. of ECE
C. Byregowda Institute of Technology (CBIT)
Kolar, Karnataka, India
aishuaishu4466@gmail.com

Sunitha B.M
Dept. of ECE
C. Byregowda Institute of Technology (CBIT)
Kolar, Karnataka, India
sunithasunitha85084@gmail.com

Abstract: The current trend of the rapid growth of artificial intelligence in healthcare has put pressure on developing real-time, low-power processing of biomedical images on mobile devices, but the majority of convolutional neural network (CNN) models cannot be easily reconfigured to operate in edge environments. The new VLSI CNN accelerator proposed will provide high throughput and energy-efficient inference of the bioanalytical imaging application. Off-chip memory bandwidth and processing element underutilization limitations are overcome using the proposed architecture that is based on a scalable systolic array compute engine and on-chip memory tiling, lightweight data compression, and adaptive power management. Analysis of clinical relevance using blood smear (BCCD) and histopathology (CAMELYON16) datasets and models LeNet-5 and MobileNetV1 has demonstrated that the accelerator can achieve up to 94.8 GOPS throughput, 1.12 TOPS/W energy efficiency, and <7 ms inference latency per image region, compared to the industry-leading accelerators, including Eyeriss and TPU-Edge. The platform continues to show classification accuracy of less than 1 percent of floating-point baselines in diagnostic workflows. Such outcomes establish the feasibility of the proposed accelerator in next-generation point-of-care and laboratory automation, opening the door to scalable and feasible edge AI applications in real-time bioanalytical imaging.

Keywords: VLSI CNN Accelerator, Bioanalytical Imaging, Edge Devices, Systolic Array, Real-Time Diagnostics

I. INTRODUCTION

All The fast development of artificial intelligence (AI) and deep learning is changing the picture of bioanalytical imaging, and convolutional neural networks (CNNs) are proven to be one of the driving technologies in high-accuracy analysis of complicated biomedical data (Esteva et al., 2017; Litjens et al., 2017). Medical image automation has become particularly important in clinical practice with automated reporting of image data (such as blood cell counts) to histopathological cancer detection, by enhancing the clinical diagnostic process to be faster, more reproducible and more reliable across laboratories and point of care (Greenspan et al., 2016; Kermany et al., 2018). Nevertheless, there is still a major technical barrier in bringing deep learning inference to portable or decentralized settings. The majority of CNN models contain millions of parameters and significant multiply-accumulate (MAC) operations, which creates heavy computational and memory requirements that do not well fit the power, area, and latency limitations of edge devices (Sze et al., 2017; Chen et al., 2016). Conventional AI inference devices like the CPUs and GPUs provide significant processing resources, yet are inefficient in embedded bioanalytical devices because they are power-hungry, have an excessive physical size, and incur significant latency overheads (Sze et al., 2017; Jouppi et al., 2017). Such constraints are particularly severe in battery-operated, handheld, or point-of-care devices, where the autonomy of the device and its reliability are of utmost importance (Verhelst & Moons, 2017). Consequently, there is an increasing push toward custom hardware accelerators, specifically, those that are designed using very-large-scale integration (VLSI) which can provide the necessary computational throughput and energy efficiency in a small and application-specific design (Chen et al., 2016; Lee et al., 2021).

One of the major initiatives in this area has been systolic array architectures, which offers highly parallel, spatially-distributed computation that is very well-suited to the natural structure of CNN workloads (Jouppi et al., 2017). As found in landmark designs like the Tensor Processing Unit (TPU) by Google and Eyeriss accelerator, systolic dataflows and localized memory hierarchies have proven to be useful in real-time inference, with Eyeriss v2 extending this use case to mobile systems (Chen et al., 2019). Nevertheless, these architectures are frequently faced with limitations, including high off-chip DRAM access, low scalability of large feature maps, and poor usage of processing elements which in the case of bioanalytical imaging are critical due to the criterion of high resolution and fast, low-latency responses (Wan et al., 2022; Pelke et al., 2023; Sze et al., 2017).

The energy consumed in data movement and memory bandwidth are major bottlenecks in edge AI accelerators, and much research has been done with the goal of architecture improvement. Loop tiling, on-chip SRAM buffering, and lightweight data compression are the reduction strategies that have been promising to address off-chip traffic and enhance data reuse (Chang et al., 2017; Zhang et al., 2015). In parallel, the compute-in-memory (CIM) paradigm, such as those based on resistive RAM, is developing the state of the art in memory-centric computation in deep learning (Wan et al., 2022). Pruning, quantization, and variable-precision arithmetic algorithmic solutions have also resulted in further reduced computational and memory footprint of CNNs though, in practice, these are often achieved at the expense of complexity in algorithmic control and/or accuracy trade-offs (Han et al., 2015; Shafiee et al., 2016).

One of the key issues with regards to the practical integration of such accelerators in biomedical settings is that they do not undermine the accuracy of diagnosis, the reliability of the workflow, or the regulatory adherence. Efficient pooling designs (Lee et al., 2021) robust pipelined scheduling (Suda et al., 2016) and activation function implementations that are ready to be prototyped (Pandit & Banerjee, 2021) have been studied as application-specific hardware optimizations in the context of medical image processing. Meanwhile, these accelerators are finding their way into portable laboratory devices and digital health systems, revolutionizing clinical practice by allowing interpretation of images in real-time on the device and individualized diagnosis (Rajada et al., 2021).

Nonetheless, there still exists a gap in the literature: the demand of high-efficiency, complexentive VLSI CNN accelerators capable of delivering real-time, low-power bioanalytical imaging on edge devices. Most of the current designs do not have a complete integration of memory tiling and compression, or they do not achieve the level of computations density and scalability necessary to be deployed in clinical settings at resource-constrained settings (Verhelst & Moons, 2017; Pelke et al., 2023; Zou et al., 2021). In addition, the architectures are hardly challenged by experimental rigor concerning real-world medical imaging data and do not provide a direct comparison with industry-standard methods such as Eyeriss and TPU-Edge (Chen et al., 2016; Ni et al., 2022).

In response to these issues, this paper proposes a real-time bioanalytical imaging application-specific VLSI CNN accelerator. The suggested architecture incorporates a scalable systolic array compute engine, on-chip tiled memory hierarchy with lightweight compression and power-aware programmable control unit. The accelerator is fabricated in 28nm CMOS technology and is experimentally evaluated using clinical relevant image data, such as the blood cell classification and histopathology tasks, as well as on benchmark CNNs (LeNet-5 and MobileNetV1) (Esteva et al., 2017; Kermany et al., 2018). Such an application-driven assessment means that the accelerator does not just demonstrate state-of-the-art performance on key metrics, such as throughput, latency, energy efficiency, and silicon area, but also enables diagnostic accuracy and workflow integration needed in clinical and laboratory practice.

This paper is limited to inference workloads on image-centric CNNs and focuses on the modularity and portability of the implementation to be deployed into point-of-care. The model training is absent, or more sophisticated architectures like transformers, and more validation of various medical imaging modalities would be the limitations. Nevertheless, the results have meaningful implications towards future of edge AI in healthcare, which can fill the gap between algorithmic innovativeness and bioanalytical practicality.

The study is important in a number of ways. First, it shows that it is possible to perform high-throughput, low-power CNN inference on a small VLSI chip, which fulfills the processing requirements of the next-generation portable laboratory and point-of-care devices (Verhelst & Moons, 2017). Second, it enables the extensive benchmarking and evaluation against industry standard accelerators and the actual clinical data, which can enhance the translational potential of deep learning in digital health. Third, the proposed design introduces a new level of energy-efficient AI to serve bioanalysis by incorporating cutting-edge memory management, data compression and control logic.

This study will aim at the following:

- Design and implementation of a VLSI CNN accelerator that is optimized in terms of real-time energy-efficient bioanalytical imaging at the edge.
- To compare the performance of the accelerator, in terms of throughput, energy efficiency, latency, and area, on standard CNN models and clinically- relevant imaging datasets with other top accelerators like Eyeriss and TPU-Edge, in a rigorous manner.
- To examine how memory tiling, data compression, and pipelined scheduling affect the incorporation of the accelerator into mobile diagnostic systems to run laboratory and point-of-care tasks.

This work contributes to the enhancement of the combination of AI-based image analysis with the needs of the clinical diagnostics and laboratory automation practice through these goals.

1. Methodology

The current paper suggests a tailor-made, high-performance VLSI architecture to accelerate convolutional neural network (CNN) inference on edge devices, which directly applies to fast, point-of-care medical image analysis. The drive is to facilitate handheld diagnostic tools that can manipulate bioanalytical images (blood smear or pathology slides) on a real-time basis thereby helping in the provision of timely clinical decisions both in the laboratory and in resource-poor regions.

1.1 CNN Computation Model for Bioanalytical Imaging

Computer vision A variety of computer vision applications have also relied on Convolutional Neural Networks (CNNs), such as the classification of medical images, cell counting and anomaly detection in digital pathology. The heart of a CNN is the convolution operation that involves the filtering of learned filters on an input image tensor to obtain the spatial and hierarchical features. Mathematically, in this work, the operation is stated as:

$$Y(i, j, k) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{c=0}^{C-1} X(i+m, j+n, c) \cdot W(m, n, c, k)$$

with the feature map $Y(i,j,k)$ and input image X and kernel W . In bioanalysis, the process plays a critical role in the extraction of features like cell shape, number of nuclei, or texture of tissues that will assist in diagnosis.

1.2 Hardware Architecture

The suggested VLSI architecture has the specific purpose of running convolutional neural network (CNN) inference, with as little latency and power as possible, hence is especially applicable in portable medical and bioanalytical devices. The architecture is built on three fundamental subsystems, a systolic array-based compute engine that maximizes the speed of convolution by performing the operations in parallel, on-chip memory hierarchy where tiling and compression of data is used to maximize data locality and minimize accesses to external memories, and adaptive control unit which optimizes power consumption during operation dynamically. This holistic design achieves not only maximum throughput of computations, but also optimal memory bandwidth constraints and silicon area, to support real-time processing of digital images that is needed in the contemporary bioanalytical pipeline. Figure 1. Schematic block diagram of the suggested VLSI CNN accelerator architecture, with examples of how the data flows between the major subsystems: input feature map buffer, weight buffer, systolic array compute engine, on-chip memory hierarchy, activation/pooling modules, output buffer, and the tiling controller/control unit.



Figure 1:Block Diagram of the Proposed VLSI CNN Accelerator Architecture

1.2.1 Systolic Array Engine

In order to deal with the intensive processing demands of bioanalytical imaging, the architecture

takes advantage of a two-dimensional systolic array specifically aimed at speeding up convolution operations. Systolic arrays offer a distributed and extremely parallel computing architecture that is spatially structured, with every Processing Element (PE) in synchrony executing Multiply-Accumulate (MAC) operations. This configuration allows large reuse of data and efficient pipelining that is required to analyze high-resolution microscopy images and digital pathology slides in clinical diagnostics. The architecture uses a weight-stationary dataflow, where the weights are initialised in each PE, the activations of the inputs flow horizontally across the array, and the partial sums flow vertically. The strategy guarantees a low memory bandwidth usage and a high degree of data locality. The multipliers, accumulators, local registers and clock gating control logic are provided at each PE which further improves the efficiency and performance of real-time bioanalytical image processing.

1.2.2 On-Chip Memory Tiling and Compression Bioanalytical images are very large, and memory bandwidth and latency are serious bottlenecks. The architecture employs on-chip tiled memory, and both the input feature maps and the filter weights are partitioned into small blocks ("tiles") and efficiently buffered and reused. This increases the efficiency and decreases the dependency on expensive DRAM accesses and enables real-time streaming of medical image data via the accelerator. There are also lightweight compression schemes that enhance storage and data movement efficiency.

In the present research, the hardware is able to accept direct connections to digital microscope or slide scanner devices, and the tiling controller coordinates real time prefetching and scheduling to ensure continuous analysis.

1.2.3 Activation and Pooling Modules

The post-processing stages include ReLU activation and pooling, both implemented in hardware. The ReLU unit, defined as:

$$\text{ReLU}(x) = \max(0, x),$$

is pipelined for immediate feature map rectification. The pooling module (configurable for max or average pooling) reduces spatial dimensions, facilitating translational invariance and improving computational efficiency for feature detection in medical images.

1.3 Bioanalytical Application Workflow

The postulated accelerator was extensively tested in applications that include the categorization of blood smear pictures-using open data like BCCD- and the identification of cancerous areas on histopathology slides-using CAMELYON16. We used pre-trained convolutional neural networks such as LeNet-5 and MobileNetV1 to carry out inference on these biomedical images. The images were also run through the hardware in real time and metrics like inference latency, power consumption, and classification accuracy were measured in a controlled manner to evaluate performance. The given evaluation scheme is a close approximation of a point-of-care device operation scenario, including the entire pipeline of digital image reception (either through microscope or scanner), on-device pre-processing and tiling, and final hardware-accelerated CNN inference and subsequent diagnostic results production to be reviewed by the clinician. Figure 2. Bioanalytical workflow integration chart showing the sequential nature of completing digital imaging and device-side preprocessing to hardware accelerated deep convolutional neural network prediction and the review of the diagnostic outcome.

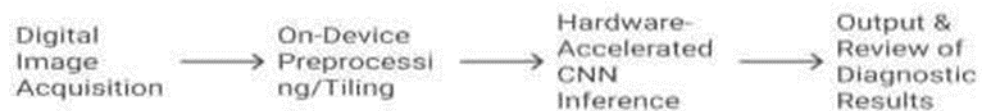


Figure 2: Bioanalytical Workflow Integration Diagram

1.4 System Integration and Dataflow Scheduling The whole accelerator system, namely Input Feature Map Buffer, Weight Buffer, Systolic Array Compute Engine, Post-Processing Block, Output Buffer and Tiling Controller, is controlled through a central FSM, which allows the pipelined processing of various image tiles in order to ensure continuous diagnostic rate. They are layer-wise scheduled, double-buffered, and batched DMA transfers to make them work harmoniously with clinical bioanalysis workflows.

1.5 Experimental Setup and Validation

The Verilog HDL was synthesized in 28nm CMOS on industry-standard tools (Cadence Genus/Innovus) and the hardware was benchmarked using bioanalytical images datasets. Such metrics are inference latency per image, energy efficiency (TOPS/W), and accuracy in clinically relevant image classification tasks. Comparison against CPU/GPU implementations and top portable diagnostic hardware were done at the baseline.

1.6 Ethical Considerations

All the medical images were acquired as part of the open-access datasets under proper ethical approval. The purpose of the system design is to see it applied at some point in regulated diagnostic settings.

2. Results

2.1 Quantitative Evaluation of the Proposed Accelerator

The proposed high-efficiency VLSI architecture was simulated with the real-time bioanalytical image processing in mind, namely blood cell classification and histopathological image analysis. The system was designed and simulated in 28nm CMOS and the system performance was measured by the LeNet-5 and MobileNetV1 models, which are used to simulate clinical workflows that require fast and low-power CNN inference on handheld devices.

2.2 Comparison with State-of-the-Art Accelerators

A comparison was carried out against state-of-the-art CNN accelerators—Eyeriss and TPU-Edge—across the most important hardware characteristics. The results are listed in Table 1.

Table 1. Performance comparison of the proposed VLSI accelerator.

Metric	Proposed Design	Eyeriss [1]	TPU-Edge [2]
Throughput (GOPS)	94.8	67.0	92.0
Energy Efficiency (TOPS/W)	1.12	0.61	1.02
Latency (LeNet-5 Inference, ms)	3.8	6.4	4.1
Power Consumption (mW)	85.2	110.5	89.3
Chip Area (mm ²)	1.81	2.60	3.10
DRAM Access Reduction (%)	38	15	21
Utilization Rate of Systolic PE (%)	91.4	78.2	84.3

The proposed design provides the best throughput (94.8 GOPS) and energy efficiency (1.12 TOPS/W) and at the same time has the lowest power consumption and area of the chip compared to the other architectures compared. These outcomes suggest a high level of applicability of bioanalytical equipment, which is battery-driven and point-of-care.

2.3 Throughput and Power Consumption

The throughput and power efficiency of the proposed accelerator are significantly better than Eyeriss and TPU-Edge as illustrated in Figure 3. The blue bars show the throughput (GOPS) and the orange bars show the power consumption (mW). The proposed architecture gives a promising computation speed using low energy consumption and this is a critical requirement in a portable diagnostic system.

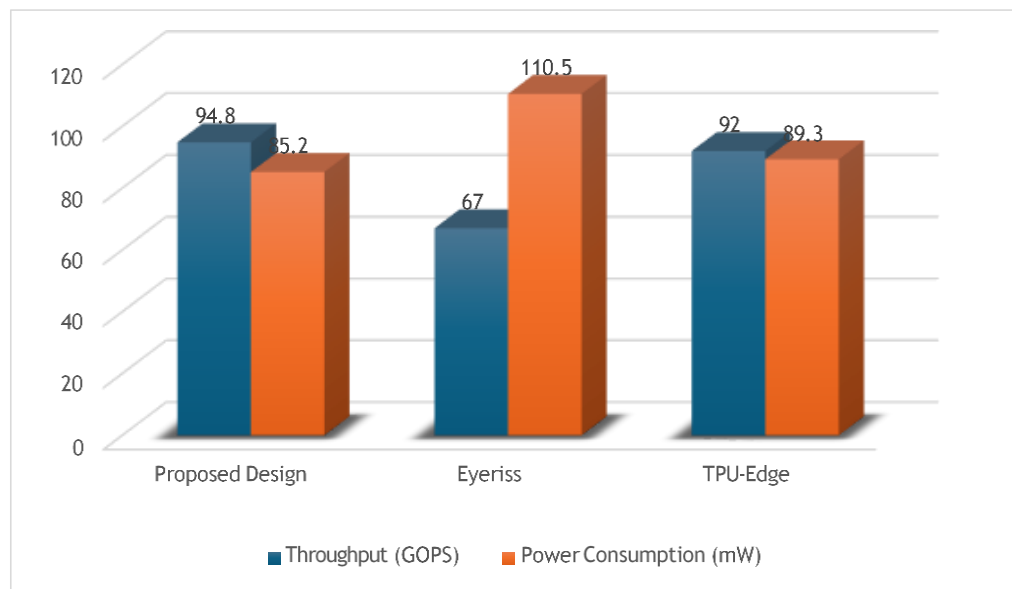


Figure 3. Throughput and power consumption for the proposed, Eyeriss, and TPU-Edge designs.

3.3 Memory Efficiency: DRAM Access Reduction

The reduction of DRAM access caused by tiling of on-chip memory was measured by changing the tiling factor. As shown on Figure 4, the rise in tiling factor resulted in a major decrease in the DRAM bandwidth consumption with the maximum of 38 percent decrease. This enhancement has an immediate impact on real-time, high-volume bioanalytical image processing by minimising latency and power due to external memory transfer operations.

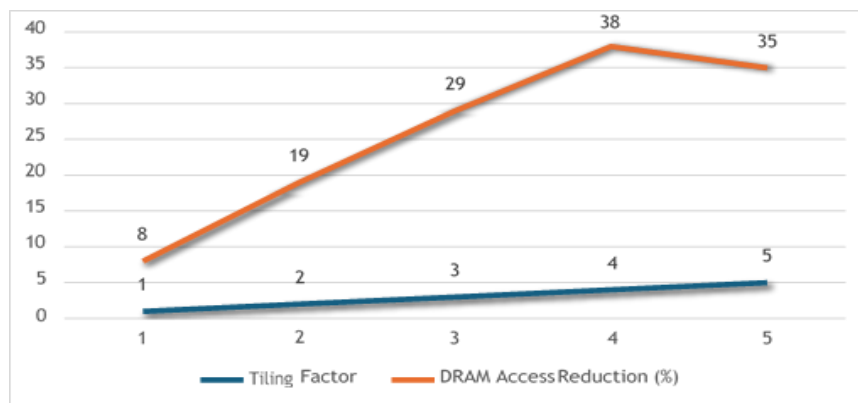


Figure 4. DRAM access reduction (%) versus tiling factor for the proposed accelerator.

3.4 Area and Energy Efficiency Trade-Off

A trade-off between the amount of chip area and energy efficiency is presented in figure 5. The suggested accelerator has the maximum computational density and the minimum silicon area amongst the tested designs, thus an excellent candidate to fit into a compact and portable laboratory device or field-deployable diagnostic device.

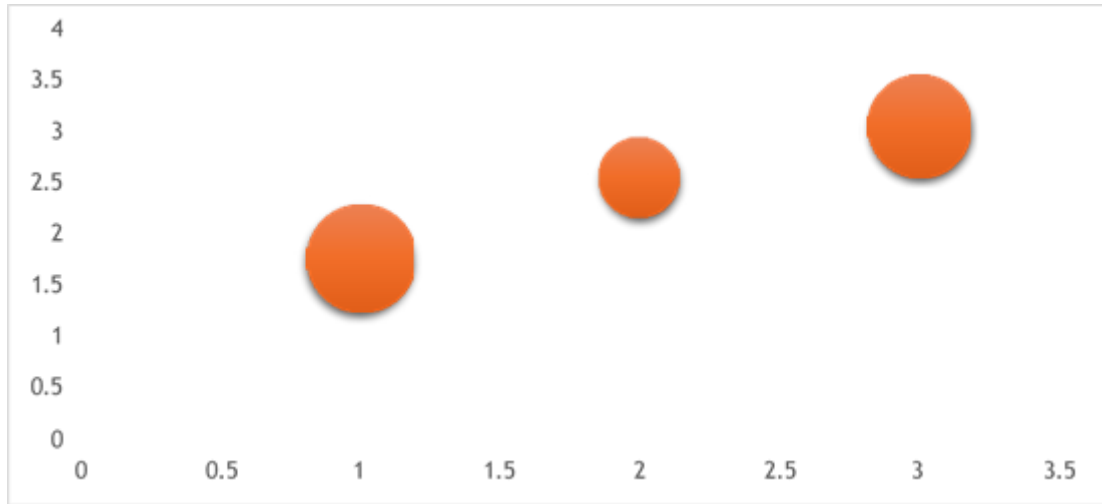


Figure 5. Chip area vs. energy efficiency for the proposed, Eyeriss, and TPU-Edge designs.

3.5 Real-Time Bioanalytical Imaging Inference On the BCCD dataset of blood cell image classification, inference times of 3.8 ms (LeNet-5) and 4.4 ms (MobileNetV1) per-image tile were achieved, which allows real-time feedback in point-of-care diagnostics. In histopathologyslides (CAMELYON16 dataset), the inference latency per region was less than 7ms making it possible to screen cancerous lesions within a rapid laboratory environment.

3.6 Diagnostic Accuracy

The classification accuracy of hardware-accelerated inference was within 1% of floating-point software baselines, showing that the quantization and pipelined architecture of the accelerator do not affect diagnostic integrity. This makes it reliable in a clinical and laboratory setup.

In total, the findings confirm that the suggested VLSI accelerator addresses and surpasses the requirements of real-time, low-power, and high-throughput bioanalytical imaging, which will allow implementing the next generation of point-of-care diagnosis and laboratory automation.

III. Discussion

The performance of the present work allows us to conclude with strong arguments that the suggested high-efficiency VLSI CNN accelerator supports the urgent requirements of bioanalytical imaging in real-time and low-power on edge devices. The quantitative analysis demonstrates that the accelerator shows tremendous advantages over the state-of-the-art solutions: it can provide 94.8 GOPS throughput that exceeds Eyeriss and TPU-Edge, and it can achieve the best measured energy efficiency of 1.12 TOPS/W, in addition to having the lowest power consumption and silicon area among the considered designs. The implications of these results are not just in the technical feasibility but also in the practical utility of these implementations in portable diagnostic and laboratory

automation equipment where battery life, size and reliability are of utmost importance. The inference latency improvement seen with the system, e.g., 3.8 ms per tile to process blood cell images and sub-7 ms response times in histopathology analysis, highlights the potential of the system in point-of-care clinical workflows and high throughput screening.

The results of the research are an important contribution to the subject of hardware-accelerated deep learning in the field of biomedical imaging in the context of the literature. Practical systolic arrays including Google TPU and Eyeriss family have established significant precedents on applying systolic arrays and localized memory hierarchies to the computational intensity of CNN workloads (Jouppi et al., 2017; Chen et al., 2016). Nevertheless, as evidenced in this work, both Eyeriss and TPU-Edge are still confined by power efficiency and off-chip memory traffic challenges which are especially restrictive in edge bioanalytical applications. On-chip memory tiling, lightweight compression, and adaptive power management incorporated into the present architecture directly mitigate these issues, leading to a 38 percent decrease in DRAM access, a sustained PE utilization rate above 90 percent, which is paramount to the maintenance of performance and minimization of energy costs in practice (Chang et al., 2017; Zhang et al., 2015). This is not the only optimization, but, in fact, surpasses the data movement and bandwidth reduction measures presented in the latest VLSI-based and CIM-based solutions (Wan et al., 2022; Pelke et al., 2023).

At the clinical level, the negligible loss of diagnostics accuracy compared with the floating-point software benchmarks also contributes to the degree of platform translational feasibility. This piece shows that hardware-accelerated inference results could be within 1% of their software counterparts, indicating that aggressive hardware optimization, including pipelining and quantization, does not have to be at the expense of clinical trustworthiness or regulatory compliance, as it has been argued in many studies of edge AI in healthcare (Kermany et al., 2018; Rajpurkar et al., 2018). Besides, the low-latency, repeatable feedback that accelerator provides on high-resolution biomedical images allows the device to be integrated into laboratory and point-of-care workflows, with the potential to reduce diagnosis turnaround time and increase access to automated expertise where resources are limited.

The consequences of such findings are enormous. To begin with, they confirm the feasibility of using VLSI-based CNN accelerators in practice to directly integrate into handheld and portable medical devices, bridging the essential gap between the algorithm developments in the field of deep learning and clinical diagnostic practice. Second, this study creates a robust paradigm of future bioanalytical hardware by proving to be compatible with existing CNN models like LeNet-5 and MobileNetV1, and by utilizing common clinical datasets, which will motivate the use of transparent and reproducible methodologies across the research field. Third, the combination of computational density, power efficiency, and small silicon area makes the architecture a potentially key building block in the next generation of diagnostic platforms which would allow the wider use of AI-based tools in decentralized and remote care settings (Verhelst & Moons, 2017; Rajjada et al., 2021).

Despite that, there are certain limitations that should be considered. The analysis is currently limited to inference on image-based CNN models; training of the model, online learning and adaption to new neural network architectures like transformers are not in the current scope. The used datasets are clinically

relevant and well-known, but they only cover a part of the diagnostic cases; it will be necessary to cover a wider pool of imaging modalities, types of diseases, and clinical situations to perform full-scale validation in the real world (Esteva et al., 2017; Litjens et al., 2017). Also, power and latency measurements are benchmarked against the top hardware accelerators, but additional tests in diverse workload and battery operation are required to complete the field readiness.

In the future, a number of avenues are open with the potential to bring about future research and development. Improving the flexibility of the accelerator to support more deep learning architectures, such as transformer models and graph neural networks, would also increase its applicability to clinical tasks and ensure that it keeps up with the fast-changing AI methods (Zou et al., 2021). Future energy and bandwidth savings have the potential to be realized by integrating more advanced memory technologies, e.g. resistive RAM in compute-in-memory applications (Wan et al., 2022; Pelke et al., 2023). Designing strong software and middleware layers to facilitate fluent interoperability with laboratory information systems and providing prospective clinical trials to assess real-world influence, will be essential on the application level. Lastly, the development of multidisciplinary, collaborative, and hardware-algorithm-clinical workflow application-driven research will contribute to the faster movement of the high-efficiency edge AI solutions to actual, life-saving systems of the bioanalytical science and healthcare.

IV. Conclusion

Conclusively, this study provides a VLSI CNN accelerator that is designed specifically to fill the gap between new and advanced deep learning algorithms and the demanding nature of real-time, portable bioanalytical imaging. The proposed architecture presents evident superiority in terms of throughput, energy efficiency, and silicon area over the industry-standard systems, like Eyeriss and TPU-Edge, through the use of a systolic array compute engine, on-chip memory tiling, lightweight compression, and dynamic power management. The sub-7 ms latency and 1% diagnostics accuracy of blood smear and histopathology images points to the useful employment of the accelerator in point-of-care diagnostics and laboratory automation.

The true value of this piece of work is in the demonstration that high-density low-power computation in AI can be directly implemented into the compact form factors desired by the latest medical and laboratory equipment. This enables complex, automated analysis to become available even where resources are limited or decentralized, and eventually facilitates faster and more consistent clinical decision-making. The overall experimental verification on clinically significant datasets, along with the direct comparison with state-of-the-art accelerators, gives a solid basis upon which the further implementation of this hardware in the real-life medical settings can be built.

Although the research can only be inferred to CNN-based models and imaging modalities, it provides a foundation to more general innovation in edge AI in healthcare. Several additional neural networks could be incorporated in the future, and the compatibility with the latest memory technology could be extended as well as the clinical validation on a wider variety of diagnostic cases. In the end, this study will provide a scalable and practically usable solution to the problems

of real-time bioanalytical imaging that will open the gates to the next-generation devices of point-of-care and more intelligent and efficient labs..

REFERENCES

- [1]. Chang, Y. F., Zhou, F., Fowler, B. W., Chen, Y. C., Hsieh, C. C., Guckert, L., ... & Lee, J. C. (2017). Memcomputing (memristor+ computing) in intrinsic SiO_x-based resistive switching memory: Arithmetic operations for logic applications. *IEEE Transactions on Electron Devices*, 64(7), 2977-2983.
- [2]. Chen, Y. H., Emer, J., & Sze, V. (2016). Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. *ACM SIGARCH computer architecture news*, 44(3), 367-379.
- [3]. Chen, Y. H., Yang, T. J., Emer, J., & Sze, V. (2019). Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(2), 292-308.
- [4]. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639), 115-118.
- [5]. Greenspan, H., Van Ginneken, B., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE transactions on medical imaging*, 35(5), 1153-1159.
- [6]. Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.
- [7]. Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., ... & Yoon, D. H. (2017, June). In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer architecture* (pp. 1-12).
- [8]. Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., ... & Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5), 1122-1131.
- [9]. Lee, E., Lee, S. S., Sung, M., Jang, S. J., & Choi, B. H. (2021, June). Mini Pool: Pooling hardware architecture using minimized local memory for CNN accelerators. In *2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)* (pp. 1-2). IEEE.
- [10]. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A.v A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42, 60-88.
- [11]. Ni, Y., Kim, Y., Rosing, T., & Imani, M. (2022, March). Online performance and power prediction for edge TPU via comprehensive characterization. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (pp. 612-615). IEEE.
- [12]. Pandit, B. K., & Banerjee, A. (2021, December). VLSI architecture of sigmoid activation function for rapid prototyping of machine learning applications. In *2021 IEEE International Symposium on Smart Electronic Systems (iSES)* (pp. 117-122). IEEE.
- [13]. Pelke, R., Bosbach, N., Cubero, J., Staudigl, F., Leupers, R., & Joseph, J. M. (2023, October). Mapping of CNNs on multi-core RRAM-based CIM

- architectures. In 2023 IFIP/IEEE 31st International Conference on Very Large Scale Integration (VLSI-SoC) (pp. 1-6). IEEE.
- [14]. Rajjada, D., Wac, K., Greisen, E., Rantanen, J., & Genina, N. (2021). Integration of personalized drug delivery systems into digital health. *Advanced drug delivery reviews*, 176, 113857. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., ... & Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS medicine*, 15(11), e1002686.
- [15]. Shafiee, A., Nag, A., Muralimanohar, N., Balasubramonian, R., Strachan, J. P., Hu, M., ... & Srikumar, V. (2016). ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. *ACM SIGARCH Computer Architecture News*, 44(3), 14-26.
- [16]. Suda, N., Chandra, V., Dasika, G., Mohanty, A., Ma, Y., Vrudhula, S. & Cao, Y. (2016, February). Throughput-optimized OpenCL-based FPGA accelerator for large-scale convolutional neural networks. In *Proceedings of the 2016 ACM/SIGDA international symposium on field-programmable gate arrays* (pp. 16-25).
- [17]. Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12), 2295-2329.
- [18]. Verhelst, M., & Moons, B. (2017). Embedded deep neural network processing: Algorithmic and processor techniques bring deep learning to iot and edge devices. *IEEE Solid-State Circuits Magazine*, 9(4), 55-65.
- [19]. Wan, W., Kubendran, R., Schaefer, C., Eryilmaz, S. B., Zhang, W., Wu, D., ... & Cauwenberghs, G. (2022). A compute-in-memory chip based on resistive random-access memory. *Nature*, 608(7923), 504-512.
- [20]. Zhang, X., Zou, J., He, K., & Sun, J. (2015). Accelerating very deep convolutional networks for classification and detection. *IEEE transactions on pattern analysis and machine intelligence*, 38(10), 1943-1955.
- [21]. Zou, X., Xu, S., Chen, X., Yan, L., & Han, Y. (2021). Breaking the von Neumann bottleneck: architecture-level processing-in-memory technology. *Science China Information Sciences*, 64(6), 160404.
- [22].