A Low Power Mixed Signal Convolutional Neural Network for Deep Learning SoC

Malik Summair Asghar[1,2], Syed Asmat Ali Shah[1,2] and Hyung Won Kim[1, a]
[1]Department of Electronic Engineering, Chungbuk National University, Cheongju, Korea.
[2]Department of Electrical and Computer Engineering, COMSATS University Islamabad, Abbottabad Campus, Pakistan.
E-mail: summair@chungbuk.ac.kr, syed@chungbuk.ac.kr, hwkim@chungbuk.ac.kr

***Abstract* - Convolutional Neural Networks (CNNs) are getting fame due to their simpler design and higher performance. However, CNNs suffer from a large area and power consumption constraints. The multiply-and-accumulate (MAC) unit, which performs the convolution operation inside a CNN, consumes a significant amount of power consumption. In this study, we propose a mixed-signal approach for implementing analog MAC unit that can replace the digital MAC unit in CNNs. The Analog MAC unit architecture is constituted from binary weighted current steering digital-to-analog (DAC) circuit and capacitors. A digital parallel interface is designed to provide input image and filter values to the MAC unit. To realize a complete CNN model a low-power analog-to-digital (ADC) is then employed at the output to convert the final value back to a digital value. When a 3×3 convolution is performed, the analog MAC unit offers a 10.7% reduction in area and a 59.2% reduction in power consumption compared to its fully digital counterparts.**

***Keywords*—Convolutional Operation, Analog Multiplier, Neural Network Accelerator, Convolutional Neural Network.**

I. INTRODUCTION

The inclusion of Artificial Intelligence (AI) in the Internet of Things (IoT) promises new horizons while presenting some challenges. Existing AI neural networks use CPU/GPU hardware architectures, which are not feasible for IoT applications with limited power. IoT applications, however, have scarce energy source and thus require low-power solutions to ensure the longevity of the devices [1].

Modern computational systems work by executing a series of instructions and therefore are not suitable for emulating massively parallel biological neural network-like systems [2]. GPUs, on the other hand, offer some degree of the parallelism and flexibility but are not optimized for power efficiency, and simulation/emulation speed is still not well suited for large number of connections of a neural network [3].

Neuromorphic architectures are designed specifically to mimic the natural biological neural network and tend to offer higher power efficiency and parallelism over the conventional CPU/GPU architectures [4]. Many neuromorphic and acceleration systems exploiting SNN [5][6][7][8] and CNN architectures [9][10][11] are presented recently.

Convolutional Neural Networks (CNNs) deliver superior performance at the expense of growing hardware costs and power consumption. CNNs realize convolutional operations using multiply-and-accumulate (MAC) units, which consume most of the power. There is a trend towards the exploration of low-power and high-performance neural processing units or neural accelerators. Recent studies have focused on developing accelerators for CNNs [12], which attempt to improve the area, power consumption and delay [13].[14] presents an architecture to achieve the inner-product (MAC) calculation by employing a sigma delta modulator (SDM) at the expense of area and power overhead. Because the recurrent use of analog-to-digital (ADCs) and digital-to-analog (DACs) converter makes them more power and area hungry to process the data.

In [15] the authors proposed an analog neuronal computation unit (ANU) for fully connected layer for a CNN model. To elevate the process, voltage, and temperature (PVT) variations the authors exploited the two pulse generators incorporated in the ReLU-embedded voltage-to-time converters (VTCs). By incorporating these additional circuits to overcome linearity issues the proposed architecture incurs complexity and additional power consumption. Some researchers are exploring mixed-signal approaches for CNNs [16][17], where some are integrating the analog compute units directly with the image sensor.

In the digital domain, the convolutional operation is performed by using multipliers and adders. For the convolution to be performed, the number of multipliers depends upon the number of filter values. Moreover, many adders are required to integrate the output of multipliers. Thus, digital MAC units occupy a huge area along with higher power consumption. This area and power constraint drifted the researcher′s interest to find the new paradigm of Analog kernels for CNN, which can not only perform convolution but can occupy very less area and consume less power.

This paper presents and implements a MAC unit for a

a. Corresponding author; hwkim@chungbuk.ac.kr

Manuscript Received Jan. 30, 2023, Revised Apr. 11, 2023 & May. 30, 2023 & Jun. 17, 2023, Accepted Jun. 19, 2023

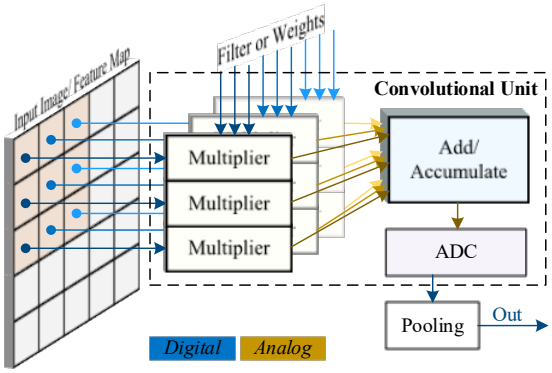

Fig. 1. Architecture of mixed-signal ACNN.

mixed-signal analog convolutional neural network (ACNN), which directly expects digital inputs for weights/filter values and image pixels. The proposed MAC unit can be adapted to use in fully connected layers as well because of the simpler and scalable design. The proposed architecture exploits the idea of binary-weighted current steering DAC and simpler current steering and accumulation circuits. Whereas the architecture implemented in [15] is fully implemented on chip starting with the on-chip image and weights memory. To convert the input image to analog format DACs are employed and then the converted analog voltage is provided

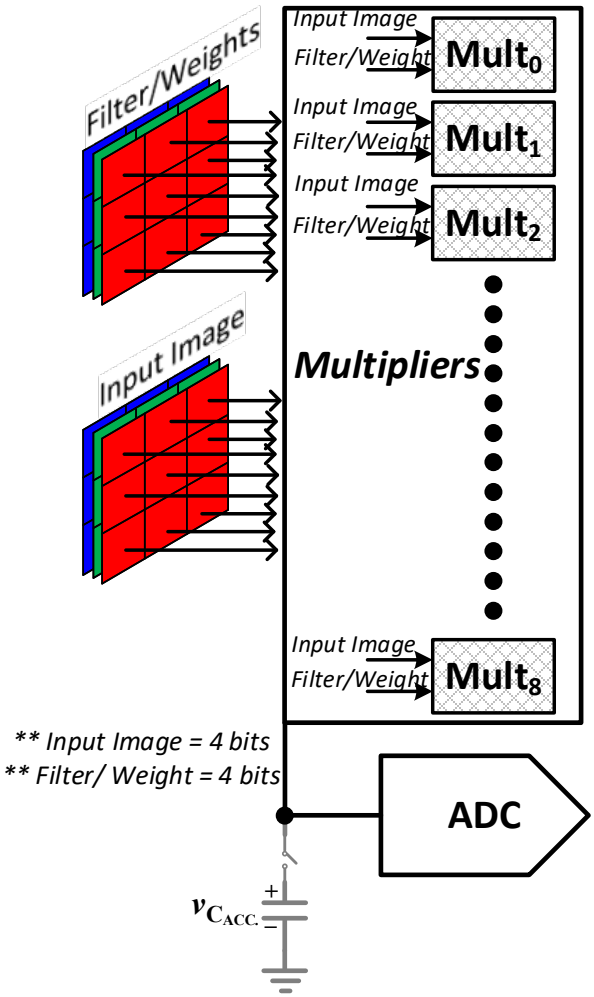

Fig. 2. Architecture of Analog Convolutional Unit.

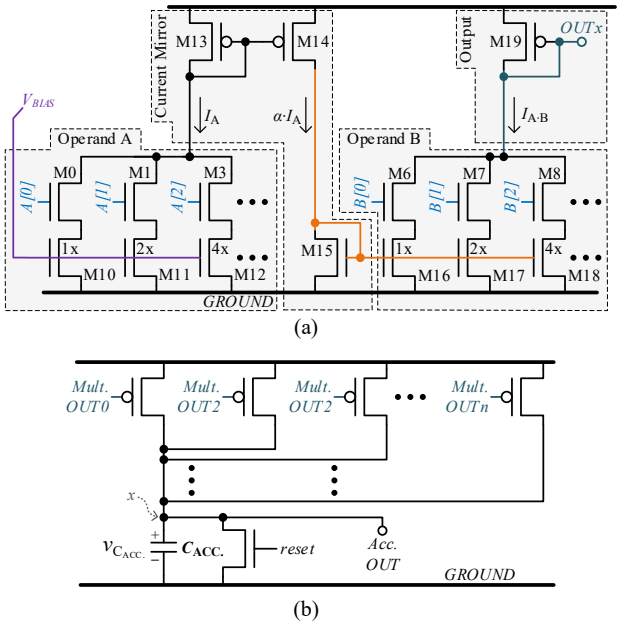

Fig. 3. Mixed-signal multiply and accumulate, a) Current steering DAC based multiplier, and b) Accumulator (current summing and integration).

to the MAC.

Section II explains the architecture of the proposed ACNN and Analog Convolutional unit (ACU). Section III describes the implementation of the individual design elements of the proposed architecture. Simulation results and performance analysis is carried out in section IV, followed by conclusion in section V.

II. ANALOG CONVOLUTIONAL NEURAL NETWORK

A. *Architecture of ACNN*

Fig. 1 shows the overall structure of mixed-signal ACNN and the use of the analog convolutional unit (CU) to replace a digital CU. The CU contains n-multipliers, for a convolutional layer with filter size $n = H_{filter} \times W_{filter}$. The output of the n-multipliers is summed together by an accumulator circuit. Finally, the output of the accumulator circuit is converted to digital domain using an ADC [14].

B. *Architecture of Analog Convolutional Unit*

Fig. 2 explains the overall architecture of the mixed-signal ACNN implementation. It consists of a single analog convolutional unit (CU) to replace the digital CU. The output of the multipliers is summed together by an accumulator circuit comprising of a configurable capacitors array. Finally, the output of the accumulator is converted to a digital value using an ADC.

III. DESIGN AND IMPLEMENTATION

A. *Multiply and Accumulate Unit.*

As the proposed Architecture aims to replace a digital CU, the inputs are digital and need to be converted to analog. Fig.

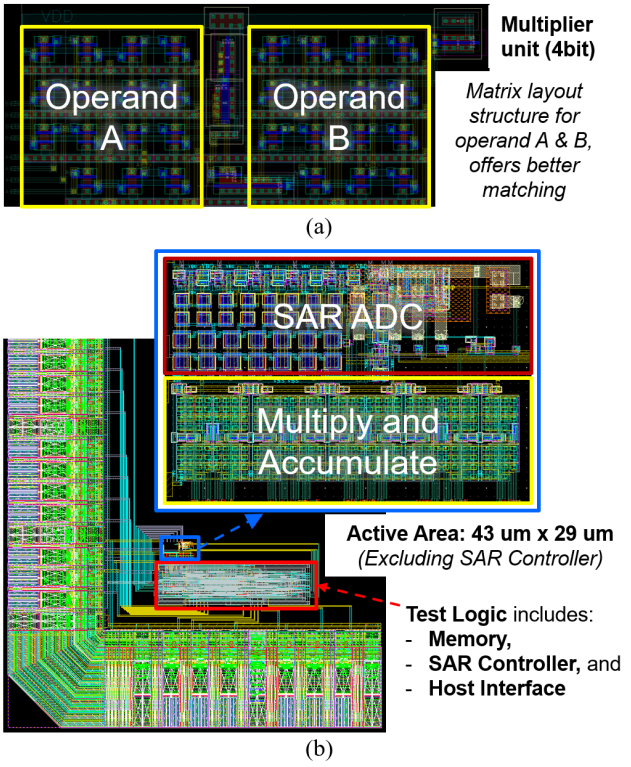

Fig. 4. The layout of: (a) Multiplier unit exploiting matrix structure of DACs (b) 3x3 analog CU with pads, ADC, and test controller.

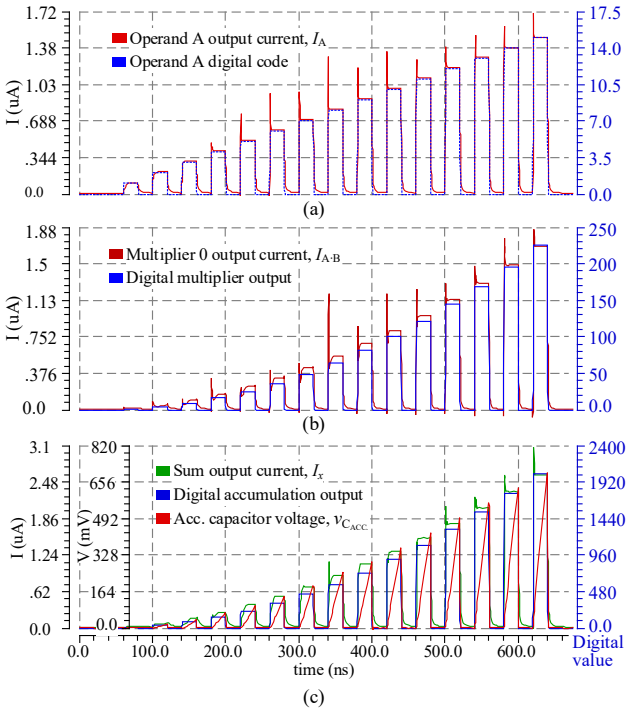

Fig. 5. Simulation results showing multiplier operand values, multiplier output current, sum output current, and accumulator output voltage.

3 (a) shows the multiplier unit with digital inputs and analog output. The multiplier tightly integrates two current steering DACs, for operands A and B. The first (left) DAC is supplied with a fixed bias voltage, and generates an output current IA, which is proportional to the digital code for operand-A. A current mirror circuit is used to generate bias voltage for the second DAC, based on the output of the first DAC. As a result, the current generated by the second (right) DAC is proportional to the product of operands A and B. Finally, the output current of the second DAC is converted to voltage using M19.

The accumulator circuit used with the multipliers is shown in Fig. 3 (b). The accumulator circuit takes output voltages from n-multiplier units and converts them into corresponding currents. These currents are summed together at node 'x', which is used to charge the accumulation capacitor C_{ACC}. Before the start of a computation, the accumulation capacitor is discharged through a MOSFET, by applying the reset signal.

B. *Implementation of ACNN*

For proof of concept, a 3x3 CU is implemented with 4-bit multipliers is implemented in CMOS 65nm library using Cadence Virtuoso design tool. A single 4-bit multiplier layout is shown in Fig. 4(a). Each of the DAC in the multiplier is constructed by placing 15-unit cells in a matrix structure, to ensure better matching compared to a simple binary-weighted DAC. The complete MAC unit, which contains nine multipliers, and one accumulate circuit, is shown in Fig. 4(b).

IV. RESULTS AND DISCUSSIONS

A. *Performance Analysis of MAC*

To verify the operation of the CU, we simulated the MAC unit (excluding the ADC), and the results are illustrated in Fig. 5. For simulations, a bias voltage of 250 mV is applied to the multipliers, while the supply is 1 V. Here, an incrementing digital code is applied, which is shared among all the operands of all the multipliers. The output current from operand-A of the first multiplier is shown in Fig. 5(a), which closely tracks the applied digital code.

Fig. 5(b) shows the output current $I_{A\cdot B}$ of the first multiplier, and the digital product of operands A and B. It can be observed that upon applying the digital code, the output current rapidly transitions to a value that closely matches the product of operands A and B.

Finally, Fig. 5(c) plots the sum of multiplier currents, observed at node 'x' of Fig. 3(b). This sum of currents charges the accumulation capacitor to reach the final value that represents the output of the MAC and matches with the digital equivalent result. The final accumulation capacitor voltage when compared to the digital MAC results, reveals a mean error of 2.83% with a peak of 5.34%. Due to the limited dynamic range of the integrated 8-bit ADC, the accumulation capacitor's voltage range is intentionally limited from 0 to 800 mV.

The simulation results for one CU are shown in Fig. 6. Here, one CU performs four iterations of MAC operation with four input images and weight values. For the first

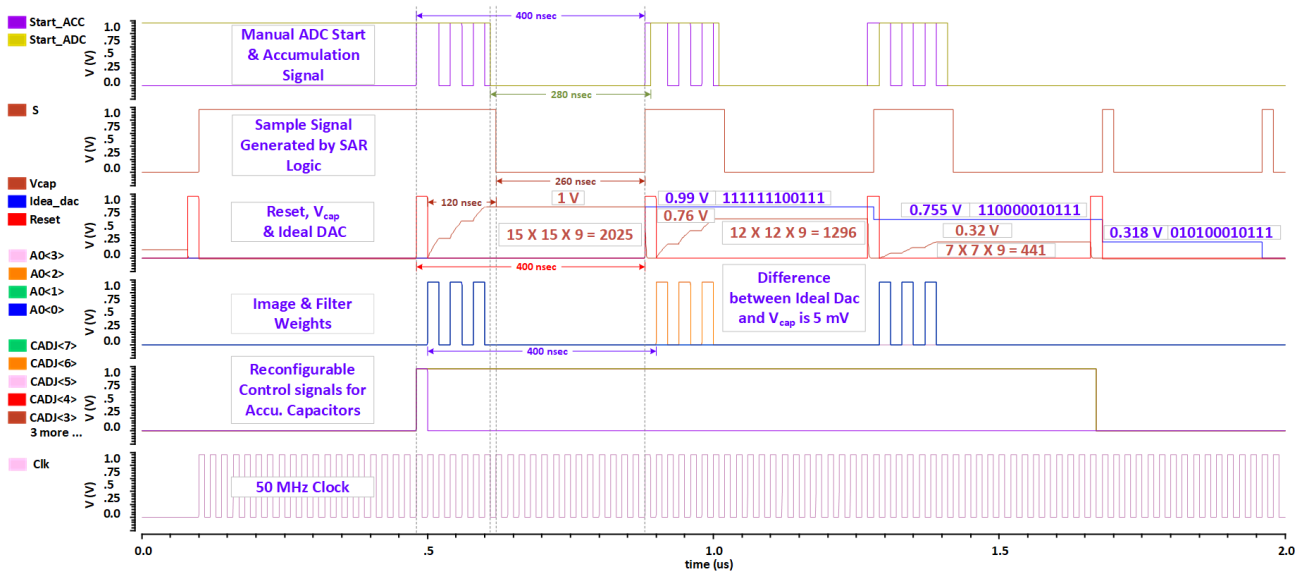

Fig. 6. Simulation results for 1 CU iterating four times.

TABLE I. Area and Power Consumption of a 3x3 Convolutional Unit (CU)

	Digital CU	Mixed-Signal CU		
		Mult. & Acc.	*ADC*	*Total*
Area	1400 μm^2	535 μm^2	715 μm^2	1250 μm^2
Power	43.11 μW	15 μW	2.6 μW	17.6 μW

iteration in the simulation, the three channels of input image and filter have a maximum value of 15 applied to the one CU. At the start accumulation signal, the first channel's value will be provided to the CU, and the equivalent voltage will be charged upon the capacitor (red waveform). The second and third channel values will be provided, respectively, and the capacitor voltage will be charged, respectively. Once accumulation is finished then, the start ADC signal will start analog to digital conversion. For verification purposes, an ideal DAC is employed to convert digital values back to analog. For the respective digital value, the ideal DAC voltage is shown in the purple waveform. Similarly, subsequent iterations will be performed.

B. Area and Power Consumption Comparison.

Table I compares the proposed and conventional (digital) 3x3 CU in terms of area and power. In addition, it lists the area and power consumption of the main blocks of the proposed CU. The equivalent digital circuit is implemented in Verilog and synthesized using Synopsys Design compiler for area and power estimates. When including the ADC, the proposed CU offers a 10.7% reduction in the area compared to the digital implementation. The proposed CU offers higher benefits in terms of power consumption, reducing it by 59.2 % when compared to the equivalent digital implementation. Benefitting from the above-mentioned area and power the proposed CU can be easily replaced for the conventional digital CUs.

Table II compares the performance of the proposed mixed-signal CU with the other state of the art analog CNN

TABLE II. Performance comparison of Convolutional Unit

Parameter	ARCHON [15]	This Work
Technology(nm)	28	65
Resolution(bits)	5(weight) 4(Input)	4(weight) 4(input)
Clock Frequency (MHz)	200	200
Area (mm^2)	0.003375	0.000535
Power(μW)	177	15
Throughput (GOPS)	.892	1.8
Energy Efficiency (TOPS/W)	5	120

processor [15]. The implementation in [15] integrates an analog neuronal computational unit (ANU) with an analog memory, on-chip registers, DACs, and ADCs. The analog data path of [15] consists of a total of 1008 processing elements. For a fair comparison the area and power consumption of [15] are normalized to a single ANU having nine MAC units. The proposed CU consisting of 9 MACs (without ADC) occupies 6x less area as compared to the one ANU of [15]. Moreover, the proposed CU (without ADC) achieves 2x times more GOPS and hence is 24x more energy efficient than [15]. The compact design of the proposed analog CU is suitable for low-power mobile applications and can be scaled easily to various CNN models.

V. CONCLUSION

This paper proposed a mixed-signal MAC unit for use as a replacement of digital MAC, in a convolutional unit (CU). The proposed CU offers promising benefits in terms of area and power consumption compared to a digital implementation. The proposed CU employs an ADC to convert accumulation results to digital. The proposed

architecture can be extended to include pooling operation in the analog domain, thereby reducing the number of ADCs and improving the speed. Moreover, the circuit can be modified to enable an analog input image directly from an image sensor, for applications where the input is acquired directly from the analog sensor.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A5A8026986) and supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-01304, Development of Self-learnable Mobile Recursive Neural Network Processor Technology). It was also supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2022-2020-0-01462) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation)" and supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1F1A1061314). In addition, this work was conducted during the research year of Chungbuk National University in 2020. The chip fabrication and EDA tool were supported by the IC Design Education Center (IDEC), Korea.

REFERENCES

[1] A. Kankanhalli, Y. Charalabidis, and S. Mellouli, "IoT and AI for Smart Government: A Research Agenda", *Government Information Quarterly*, vol. 36, no. 2, pp. 304-309, 2019.

[2] R. A. Nawrocki, R. M. Voyles and S. E. Shaheen, "A Mini Review of Neuromorphic Architectures and Implementations", *IEEE Transactions on Electron Devices*, vol. 63, no. 10, pp. 3819-3829, Oct. 2016.

[3] J.C. Knight, and T. Nowotny, "GPUs Outperform Current HPC and Neuromorphic Solutions in Terms of Speed and Energy When Simulating a Highly-Connected Cortical Model", *Frontiers in Neuroscience*, vol. 12, pp. 941, 2018.

[4] S. Gupta, "Neuromorphic Hardware: Trying to Put Brain into Chips", Jun. 30, 2019. [Online]. Available: https://towardsdatascience.com/neuromorphic-hardware-trying-to-put-brain-into-chips-222132f7e4de [Accessed on 2020-02-06 at 12:47 (GMT+9)].

[5] H. Kim, S. Hwang, J. Park, S. Yun, J. Lee and B. Park, "Spiking Neural Network Using Synaptic Transistors and Neuron Circuits for Pattern Recognition with Noisy Images", *IEEE Electron Device Letters*, vol. 39, no. 4, pp. 630-633, April 2018.

[6] H. Tang, H. Kim, H. Kim and J. Park, "Spike Counts Based Low Complexity SNN Architecture with Binary Synapse", *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 6, pp. 1664-1677, Dec. 2019.

[7] G. K. Chen, R. Kumar, H. E. Sumbul, P. C. Knag and R. K. Krishnamurthy, "A 4096-Neuron 1M-Synapse 3.8-pJ/SOP Spiking Neural Network with On-Chip STDP Learning and Sparse Weights in 10-nm FinFET CMOS", *IEEE Journal of Solid-State Circuits*, vol. 54, no. 4, pp. 992-1002, April 2019.

[8] M. S. Asghar, S. Arslan, and H. Kim, ″A Low-Power Spiking Neural Network Chip Based on a Compact LIF Neuron and Binary Exponential Charge Injector Synapse Circuits″, *Sensors*, vol. 21, no. 13, p. 4462, Jun. 2021, doi: 10.3390/s21134462.

[9] Y. A. Bachtiar and T. Adiono, "Convolutional Neural Network and Maxpooling Architecture on Zynq SoC FPGA", *International Symposium on Electronics and Smart Devices (ISESD)*, Badung-Bali, Indonesia, 2019, pp. 1-5.

[10] S. Sabogal, A. George and G. Crum, "ReCoN: A Reconfigurable CNN Acceleration Framework for Hybrid Semantic Segmentation on Hybrid SoCs for Space Applications", *IEEE Space Computing Conference (SCC)*, Pasadena, CA, USA, 2019, pp. 41-52.

[11] Y. Halawani, B. Mohammad, M. Abu Lebdeh, M. Al-Qutayri and S. F. Al-Sarawi, "ReRAM-Based In-Memory Computing for Search Engine and Neural Network Applications", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 2, pp. 388-397, June 2019.

[12] S.-S. Park and K.-S. Chung, ″CENNA: Cost-Effective Neural Network Accelerator″, *MDPI Electronics*, vol. 9, no. 1, p. 134, Jan. 2020.

[13] H. Kwon, A. Samajdar, and T. Krishna, ″MAERI: Enabling Flexible Dataflow Mapping over DNN Accelerators via Reconfigurable Interconnects″, *SIGPLAN Not.*, vol. 53, no. 2, pp. 461–475, Nov. 2018.

[14] F. N. Buhler et al., ″A 16-channel Noise-Shaping Machine Learning Analog-Digital Interface,″ IEEE Symp. VLSI Circuits, 2016.

[15] J. -O. Seo, M. Seok and S. Cho, "ARCHON: A 332.7TOPS/W 5b Variation-Tolerant Analog CNN Processor Featuring Analog Neuronal Computation Unit and Analog Memory," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 258-260, doi: 10.1109/ISSCC42614.2022.9731654.

[16] M. S. Asghar, M. Junaid, H. W. Kim, S. Arslan and S. A. Ali Shah, "A Digitally Controlled Analog kernel for Convolutional Neural Networks", *2021 18th International SoC Design Conference (ISOCC)*, Jeju Island, Korea, Republic of, 2021, pp. 242-243, doi: 10.1109/ISOCC53507.2021.9613851.

[17] J. Zhu, Y. Huang, Z. Yang, X. Tang and T. T. Ye, "Analog Implementation of Reconfigurable Convolutional Neural Network Kernels", *2019 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Bangkok, Thailand, 2019, pp. 265-268, doi: 10.1109/APCCAS47518.2019.8953177.

Malik Summair Asghar received his B.S. degree in electronics engineering from COMSATS University Islamabad, Abbottabad Campus, Pakistan, in 2009. He received his M.S. degree in electrical engineering with specialization in communication electronics from Linkoping University, Linkoping, Sweden, in 2013. He is currently working as Ph.D. student in MSIS lab at Chungbuk National University, Cheongju, South Korea. His research interests are in the areas of analog and mixed-signal circuits IC Design for AI applications, Analog front ends for touchscreen panels and design of injection locked frequency dividers.

Syed Asmat Ali Shah received his B.S. in Computer Engineering from COMSATS Institute of Information Technology (CIIT), Abbottabad, Pakistan, in 2007, M.S. in System-on-Chip from Linkoping University, Sweden, in 2012, and PhD in Electronics Engineering department from Chungbuk National University, in 2020. He served as an assistant professor at COMSATS University Islamabad, Abbottabad Campus until 2016. He is currently a post-doctoral candidate at Electronics Engineering Department, Chungbuk National University, South Korea. His current research interests include ultra-low-power circuits, power management circuits, analog, and mixed-signal Convolutional Neural networks, and SoC design.

HyungWon Kim received the B.S. and M.S. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology in 1991 and 1993, respectively, and the Ph.D. degree in electrical engineering and computer science from the University of Michigan, Ann Arbor, MI, USA, in 1999. In 1999, he joined Synopsys Inc., Mountain View, CA, USA, where he developed electronic design automation software. In 2001, he joined Broadcom Inc., San Jose, CA, USA, where he developed various network chips, including a WiFi gateway router chip, a network processor for 3G, and 10 gigabit ethernet chips. In 2005, he founded Xronet, Inc., a Korea-based wireless chip maker, where he managed the company as CEO to successfully develop and commercialize wireless baseband and RF chips and software, including WiMAX chips supporting IEEE802.16e and WiFi chips supporting IEEE802.11a/b/g/n. Since 2013, he has been with Chungbuk National University, Cheongju, South Korea, where he is currently an Associate Professor with the Department of Electronics Engineering. His current research focuses cover the areas of neural network processor SoCs, CNN optimization, object recognition for autonomous driving, V2X network and security, sensor read-out circuits, touch screen controller SoC, and wireless sensor networks.