Implementing an Integrated Neural Network for Real-Time Position Reconstruction in Emission Tomography With Monolithic Scintillators

Di Giacomo, S.; Ronchi, M.; Borghi, G.; Schaart, D. R.; Carminati, M.; Fiorini, C.

DOI
10.1109/TRPMS.2024.3378421

Publication date
2024

Document Version
Final published version

Published in
IEEE Transactions on Radiation and Plasma Medical Sciences

Citation (APA)
Di Giacomo, S., Ronchi, M., Borghi, G., Schaart, D. R., Carminati, M., & Fiorini, C. (2024). Implementing an Integrated Neural Network for Real-Time Position Reconstruction in Emission Tomography With Monolithic Scintillators. *IEEE Transactions on Radiation and Plasma Medical Sciences*, *8*(5), 501-510. https://doi.org/10.1109/TRPMS.2024.3378421

Implementing an Integrated Neural Network for Real-Time Position Reconstruction in Emission Tomography With Monolithic Scintillators

S. Di Giacomo, *Graduate Student Member, IEEE*, M. Ronchi, *Student Member, IEEE*, G. Borghi, *Member, IEEE*, D. R. Schaart, *Senior Member, IEEE*, M. Carminati, *Senior Member, IEEE*, and C. Fiorini, *Senior Member, IEEE*

***Abstract*—Embedding signal processing in the front-end of radiation detectors represents an approach to cope with the growing complexity of nuclear imaging scanners with increasing field of view (i.e., higher number of channels). Machine learning (ML) offers a good compromise between intrinsic image reconstruction performance and computational power. While most hardware accelerators for ML are based on digital circuits and, thus, require the analog-to-digital conversion of all individual signals from photodetectors, an analog approach allows to streamline the pipeline. We present the study of an analog accelerator implementing a neural network (NN) with 42 neurons in a 0.35-μm CMOS process node. The specific target is the reconstruction of the position of interaction of gamma-rays in the scintillator crystal of Anger cameras used for PET and SPECT. This chip can be used stand-alone or monolithically integrated within the application specific integrated circuit (ASIC) for the filtering of current signals from arrays of silicon photomultipliers (SiPMs). Computation is performed in charge domain by means of crossbar arrays of programmable capacitor. The architecture of the 64-input ASIC and the training of the NN are presented, discussing the impact of weight quantization on 5 bits. From MATLAB and circuit simulations, consistent with ASIC topology and operations, the NN capabilities were tested using two different datasets, obtained from both simulated data and experimental data, both based on PET detector composed by a monolithic scintillator crystal readout by an 8×8 array of SiPMs. Simulations show an achievable spatial resolution better than 2-mm full-width-at-half-maximum with a 10-mm thick crystal, a max. count rate of 200 kHz and the energy efficiency per inference is estimated to be of 93.5 GOP/J, i.e., competitive with digital counterparts, with an energy consumption of 38 nJ per inference and area of 23 mm^2.**

***Index Terms*—Analog neural network (NN), anger camera, in-memory computing (IMC), PET, position sensitivity, SPECT.**

I. Introduction

One of the prominent quests in the field of nuclear medical imaging is the continuous improvement of detector sensitivity and intrinsic spatial resolution, which result in higher quality images [1]. These often diverging requirements are combined with the need for more compact front-end circuits and a decrease of latency, power consumption and cost. Gamma-ray detectors are the fundamental technology for emission tomography instrumentation (such as PET and SPECT scanners). Their development is facing the challenges of increasing the number and density of detectors, while still preserving low power consumption and good performance. For example, an emerging trend is represented by high-sensitivity systems such as PET scanners with long axial field of view (LA-FOV) covering the majority of the body (total body PET) [2], [3]. The advantage of these systems is a superior image quality and/or dose reduction to the patient, at the price of a very large number of detector signals to be readout, increasing hardware complexity and costs.

The replacement of conventional bulky photomultiplier tubes (PMTs) with miniaturized solid-state silicon photomultipliers (SiPMs) was the recent, most significant development characterizing the current generation of state-of-the-art SPECT and PET systems. This development led to MRI compatibility of PET/SPECT scanners [4] and to a significant improvement in their energy and timing resolution, determining an extensive commercial introduction of time-of-flight (TOF) PET scanners [5], [6]. Systems based on array of SiPMs coupled to pixelated scintillator crystal of approximately 3 mm × 3 mm × 20 mm are the state-of-the-art clinical detectors, resulting in spatial resolution in the range 3–5-mm full-width-at-half-maximum (FWHM) at the system level [7].

New detectors based on monolithic scintillators have recently proven that they can achieve higher detection efficiency, with spatial resolution better than 2-mm FWHM and providing depth-of-interaction (DOI) information. The 3-D position of interaction of the impinging gamma rays in the crystal is derived from the shape of the scintillation light distribution by means of position estimation algorithms [8]. Typically, the estimate of both 2-D spatial resolution and DOI requires long and meticulous calibration procedures based on a fine sampling of the detector response over a huge number of irradiation positions in coincidence with a reference

Manuscript received 7 September 2023; revised 14 February 2024; accepted 12 March 2024. Date of publication 18 March 2024; date of current version 3 May 2024. This work was supported by the Politecnico di Milano. *(Corresponding author: S. Di Giacomo.)*

This work did not involve human subjects or animals in its research.

S. Di Giacomo, M. Ronchi, G. Borghi, M. Carminati, and C. Fiorini are with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy, and also with the Istituto Nazionale di Fisica Nucleare, Sezione di Milano, 20133 Milan, Italy (e-mail: susanna.digiacomo@polimi.it).

D. R. Schaart is with the Department of Radiation Science and Technology, Delft University of Technology, 2629 JB Delft, The Netherlands.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TRPMS.2024.3378421.

Digital Object Identifier 10.1109/TRPMS.2024.3378421

detector [9]. There are commercially available preclinical and clinical PET systems based on monolithic scintillators. One of the first preclinical scanners was the Albira system [10] based on 48 mm × 48 mm × 10 mm LYSO scintillators coupled to 12 × 12 SiPM arrays, and resulting in a spatial resolution of 0.7-mm FWHM and a DOI resolution of 2.5-mm FWHM. The MOLECUBES β-CUBE scanner [11] is a preclinical PET system composed of 25 mm × 25 mm × 8 mm LYSO crystals readout by SiPM arrays, achieving a spatial resolution of 0.84-mm FWHM and a DOI of 1.6-mm FWHM. The company Oncovision commercializes a brain PET system named CareMiBrain [12] based on 50 mm × 50 mm × 15 mm LYSO crystals coupled to 12 × 12 SiPM arrays, with a spatial resolution of 1.4-mm FWHM and a DOI better than 3-mm FWHM. There are also academic PET systems such as the MINDView brain PET insert [13], designed with LYSO crystals of 50 mm × 50 mm × 20 mm and achieving a spatial resolution of 1.6-mm FWHM and a DOI less than 4-mm FWHM.

Monolithic scintillators are gaining more and more interest with respect to pixelated crystals, thanks to their superior spatial resolution which is not limited by the pixel size, higher sensitivity due to a more efficient light transport, possibility to measure DOI and TOF capabilities. The major challenges associated with monolithic crystals are the requirement for complex and time-consuming calibration procedures due to variation in crystal and photodetector properties, truncation of scintillation light at the crystal edges due to reflections at the boundaries, loss of photon multi-interaction information owing to unavoidable light sharing among SiPMs [14]. These problems can be addressed by employing different crystal treatments and configurations or by characterizing the scintillation light distribution profiles with mathematical models, such as Lambertian light distribution [15] or the inverse square law [16], to cite a few.

As an alternative to monolithic crystals, pseudo-monolithic crystals, also called slabs, have been proposed with the aim to combine the advantages of both segmented and monolithic-based detectors. They are able to provide DOI information, and the density of scintillation photons is higher compared to monolithic crystals. Moreover, only a fraction of readout channels is involved in each interaction event, and are potentially cheaper than or competitive with large monolithic blocks. However, slabs require calibration procedures and thus share the related challenges of monolithic-based detectors. Examples of slab-based detectors reported in the literature are [17], [18], and [19]. Besides, the high sensitivity molecular imaging (IMAS) project [20] aims at realizing a total-body PET system with DOI and TOF capabilities using semi-monolithic crystals and a signal-reduction readout.

Besides the technological and instrumentation development, much progress has been made in the advancement of positioning algorithms to estimate the gamma-ray interaction coordinates. Examples of statistical methods are least square [16], [17], [21], maximum likelihood [22], [23], and k-nearest neighbors [9], [24]. Many research groups are putting a lot of effort toward the integration of artificial intelligence (AI) into medical imaging [25], [26]. Methods based on machine learning (ML) are more robust and provide high degree of generalization, being able to handle a huge amount of data. Examples are gradient tree boosting [27], [28] and neural networks (NNs) [29], [30], [31], [32], [33], [34].

However, all these approaches are power-hungry, computationally demanding and require to handle large volumes of data, since the data acquisition system (DAQ) has to record the whole light distribution seen by each SiPM matrix for every detected gamma-ray photon. Thus, the overall data bandwidth and the number of analog-to-digital conversion (ADC) channels are huge. To give an idea of the order of magnitude, in a PET scanner, one can consider a maximum coincidence-event acquisition rate in the order of several kcps up to few tens of Mcps, depending from the scanner geometry and axial length [35]. Each coincidence involves two modules, which in the case of monolithic detectors have several tens of readout SiPMs each. Since each SiPM produces approximately 4 bytes per event (for time, charge, and position), a maximum data rate in the order of several gigabytes per second can be expected. Dedicated readout electronics are used to digitize and process the photodetector signals, such as graphic processing units (GPUs) or application-specific integrated circuits (ASICs) combined with field-programmable gate array (FPGA). Requirements in terms of power budget and data bandwidth of standard architectures put a limit for fully exploiting the benefits of novel artificial intelligent-based algorithms for position reconstruction in practical implementations. However, recent increase in computing performance is enabling near real-time processing even with standard computers. For example, Benjamin et al. [36] presented a software-based architecture that allows real-time acquisition and processing of a preclinical PET/MRI-insert.

The goal of this work is to study and develop a new concept of gamma-ray detector, where the processing of the scintillation light is moved from the back-end to the front-end, allowing for an embedded computation near the photosensors. This is enabled by the design of an ASIC implementing, in silicon, an analog NN for position-of-interaction reconstruction of the scintillation event. With this ASIC we address the limitation of the current technology by providing, at the same time, improved accuracy, reduced power consumption and lower cost. The proposed ASIC is intended to be inserted in the analog acquisition chain of a typical gamma camera based on a monolithic scintillator readout by an SiPM matrix, as shown in Fig. 1, with the task of processing the analog signals coming from the SiPMs and producing as an output the interaction coordinates of the gamma photon absorbed in the scintillator crystal. This would drastically reduce the required interconnections and eliminate the need of components, such as ADCs and FPGAs for position sensitivity purposes and, eventually, the estimate of other interaction parameters (time, energy, and DOI), resulting in less power consumption, space, complexity, and cost.

In Section II, we provide a brief description of the context and background to which this work belongs, followed by a detailed illustration of the proposed analog NN architecture and its fundamental operations to perform an inference. Section III illustrates the preliminary modeling of the NN,

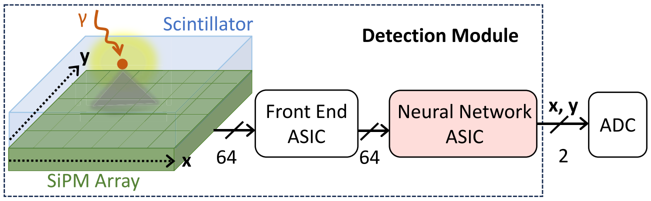

Fig. 1. Gamma camera featuring a monolithic scintillator readout by an array of SiPM detectors, and the proposed analog NN ASIC for positioning of the scintillation event.

its validation with both a simulated dataset and an experimental dataset acquired with a state-of-the-art monolithic PET detector, followed by Section IV in which a brief analysis of the electronic circuit non idealities and how they affect the network positioning performances is reported. Finally, a description of the ASIC structure and its expected energy efficiency is provided.

II. ASIC Concept and Architecture

A. In-Memory Computing

NNs are increasingly gaining a lot of interest in the medical imaging field, since they have proven to achieve very fast positioning estimation with superior spatial resolution, and to be able to accurately model the photodetector response, also taking nonlinearities into account, therefore reducing positioning bias at the edges of the detector where linear estimation methods fail [31], [32], [33], [34], [37], [38], [39], [40].

Conventional hardware platforms for training and inference of AI models are CPUs and GPUs [41]. These general-purpose processors are the state-of-the-art architectures for AI algorithms, but they suffer several limitations in terms of computational resources and memory bandwidth, especially when large volume of data and complex models have to be handled. To accelerate the inference process, improve efficiency and achieve better performance, several companies have proposed their custom hardware accelerators [42], [43], [44], in the form of ASICs, such as Google's tensor processing unit (TPU) [45], IBM True North [46] and Mythic Intelligent Processing Unit accelerators [47], [48], as well as FPGAs, such as Microsoft Brainware [49] and Intel Arria [50]. ASICs can achieve very high energy efficiency (in the order of TOP/s/W) thanks to the highly specific computing task and data paths; however ASICs have low flexibility and cannot adapt rapidly to the development of new AI algorithms. In contrast, FPGAs are more customizable and can be reconfigured to implement different algorithms. They can achieve energy efficiency of several GOP/s/W [51], [52].

Compared to traditional digital processors, the peculiarity of the proposed electronic circuit is to execute NN computations in an fully analog way within the memory cells. This chip falls in the category of AI accelerators, where computations are performed directly inside the memory, differently from the traditional Von Neumann architecture which is based on separate processing unit and memory for instructions and data storage. The frequent data movement on a single central bus between processor and memory limits the performance and energy efficiency of Von Neumann-based platforms. Given the limited bandwidth of the bus, the processor must separately access the code and data sequentially, experiencing the "Von Neumann bottleneck" [53]. A different hardware approach in which the memory is employed for both storage and computation could produce significant advantages in terms of amount of transmitted data, latency, and energy efficiency. This novel computing architecture is expected to overcome the limitations of the memory wall (disparity in speed between memory and processing units) by co-locating computational tasks with memory itself organized as a crossbar array of memory elements. This approach is referred to as in-memory computing (IMC) and enables efficient parallel computing with negligible data movement [54], [55], [56], [57].

The idea that implementation of NNs into dedicated hardware where multiple computational elements are connected in parallel has been supported by researches since the beginning of 1980s [58]. Some of early analog implementations of NNs were intended to be used for pattern classification applications (image detection, speech recognition, etc.) and use resistors [59], charge-coupled devices [60], capacitors [61], [62], and floating gate EEPROMS [63] as memory elements to store network weights. However, until the early 1990s, NN hardware has experienced a slow and modest development and commercialization, mostly due to the rapid growth of computing power of conventional general purpose processors (CPUs, DSPs, etc.) [64]. From the last two decades, a new interest was given for NN accelerators. Various approaches have been proposed, in the form of FPGAs [65], ASICs [66], GPUs and TPUs [67], [68], fully analog implementations [69], [70], and several surveys have been made [42], [43], [71], [72], [73].

We propose an analog accelerator that uses capacitors as memory elements and performs operations in the charge domain. To our knowledge, this is the first study implementing an analog NN for position sensitivity applications in the area of PET and SPECT.

B. Analog Neural Network Architecture

The fundamental operation at the base of an NN (and in general of any ML algorithm that exploits vector–matrix multiplications) is the multiply-and-accumulate (MAC) operation, where each layer input is multiplied by a weight and summed to the other weighted inputs to produce an output, which in turn is fed to the neurons of the following hidden or output layers. A nonlinear activation function is applied to each weighted sum, and its nonlinearity is fundamental to model very complex nonlinear systems.

The proposed architecture of the analog NN is a crossbar array of programmable switched capacitors that performs the MAC operations in charge domain. The analogy between a feedforward NN structure and its analog implementation is shown in Fig. 2. Specifically, the input photodetector voltages are converted in charge into programmable capacitors (of which the capacitance value is the weight) for multiplication. All the weighted input charges are then summed column-wise thanks to the virtual ground of a charge integrator, that converts the total charge into the output voltage (neuron output) by means of the feedback capacitor C_F. The activation function

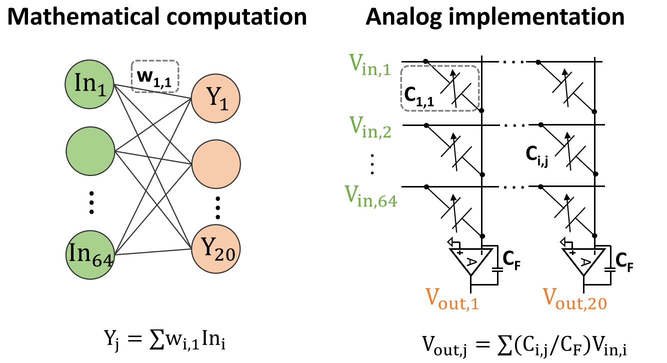

Fig. 2. Analogy between the mathematical computation of the MAC operation and its implementation with a programmable capacitive crossbar array in analog domain. The neuron output is a weighted sum of the input signals.

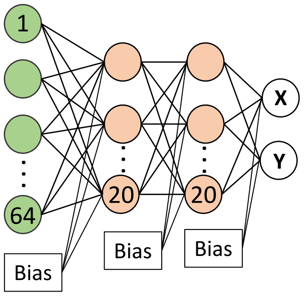

Fig. 3. NN architecture with 64 inputs (corresponding to the flattened output signals coming from an 8×8 array of SiPMs), 2 hidden layers of 20 fully connected neurons, and 2 output voltages that encode the *x* and *y* interaction coordinates of the gamma radiation.

is a clipped ReLU directly implemented by the response of the integrator itself, clipping the output voltage between the two power supplies.

In this study, we take as reference architecture an NN with 64 inputs (corresponding to the flattened output signals coming from an 8×8 array of SiPMs), 2 hidden layers of 20 fully connected neurons, and 2 output voltages that encode the *x* and *y* interaction coordinates of the gamma photon, as shown in Fig. 3. The two coordinates can assume continuous values in the range ± 25 mm, which are converted in continuous voltage values between 0 and 3.3 V. The hyperparameters of this NN were optimized by means of Grid Search Cross-Validation technique, and the final chosen architecture is the best tradeoff between network complexity and positioning performances. In order to implement the bias (the constant value which is added to the weighted product), each network layer has an additional input that is summed to each neuron of the following layer. The use of an NN allows to significantly reduce the number of signals to be fed out of the detection modules (in this specific application they are reduced from 64 to 2) with an evident impact in streamlining the interconnections (cables, connectors) and in the acquisition section. The removal of ADCs for each pixel in the detector array together with the reduction of processing operations in the FPGA allow to reduce power and speed up the computations.

C. Analog Neuron Operation

The neuron, in the classical perceptron form, is the core processing block that performs the MAC operation, and its analog implementation as a switched capacitor circuit is shown in Fig. 4. It is repeated for every neuron in every layer of the NN, with its output connected to the inputs of the neurons of the following layer. The programmable capacitor $C_{i,j}$ represents the network weight and it is made of a bank of capacitors (scaled as power of 2) selectively connected in parallel by digitally controlled switches, able to implement weights with 4-bit resolution. The impact of the quantization of weights was studied and the number of 5 bits (4 + 1 for the sign) resulted as an optimal compromise between achievable spatial resolution and neuron complexity. We also took into consideration the chip area, since the majority of the ASIC surface is occupied by capacitors, i.e., by weights, and more bits means more capacitors to be employed. Quantization with a number of bits equal to 4 would degrade the network performance of about 35%, instead using 3 bits the performance worsens of about 120%. Moreover, when using a number of bits less than 5, spatial resolution metrics, such as FWHM, are no more reliable since the PSFs of the error along *x* and *y* directions do not have a Gaussian shape anymore. The performance turns out to be weakly sensitive on the quantization of weights when 5-bit quantization-aware training (QAT) is adopted.

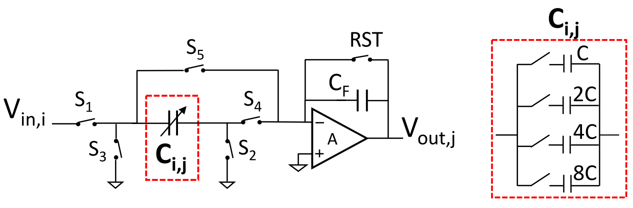

Fig. 4. Schematic of a single-input–single-output analog neuron composed by 4-bit programmable switched-capacitor bank (network weight) and charge integrator.

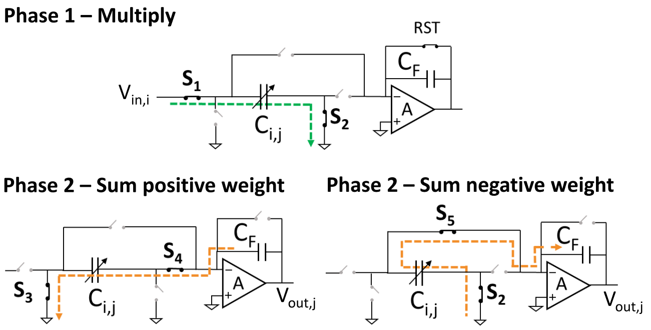

Fig. 5. Schematic illustration of the analog neuron phases needed to execute the multiply–accumulate operation, implementing both positive and negative weights.

Another set of switches (from S_1 to S_5) allows to execute the multiply–accumulate operation. A detailed illustration of how the analog neuron performs multiplication and addition is shown in Fig. 5. Initially, in phase 1, switches S_1 and S_2 are closed to charge the weighting capacitor $C_{i,j}$ (whose weight value is set by the 4 bits shown in Fig. 4) with the input voltage $V_{\text{in},i}$. Next, in phase 2, switches S_1 and S_2 are open while S_3 and S_4 are closed to connect the capacitor to the integrator virtual ground and transfer its charge to the feedback capacitor C_F. If a negative weight has to be implemented, the weighting capacitor is flipped by closing, instead, switches S_2 and S_5, thus subtracting charges from the integrator. This extra sign bit allows to achieve a 5-bit precision of the MAC operation. The integrator output is equal to $V_{\mathrm{out},j} = \sum_{i=1}^{N} V_{\mathrm{in},i}(C_{i,j}/C_F)$, where N is the total number of inputs, given by the contribution of all weighted inputs connected to the same integrator. The bias of each neuron is implemented as a fixed input signal and weighted as the other inputs.

The amount of energy lost in charging operations is crucial for the power dissipation performance of the circuit and must be minimized. In order to minimize such charging energy, a different weighting sequence has been adopted, based on charge redistribution: rather than charging $C_{i,j}$, the input voltage is sampled only on the LSB capacitor (i.e., C in Fig. 4), and successively the stored charge is shared among all the other capacitors connected in parallel. By doing this, the required energy is reduced by a factor of 15. To implement the product by the weight, some capacitors are disconnected before the summing phase, and only the remaining capacitors are connected to the integrator virtual ground to integrate a fraction of the previously stored charge. In this case, the integrator output is equal to $V_{\mathrm{out},j} = \sum_{i=1}^{N} V_{\mathrm{in},i}(C_{i,j}/15C_F)$.

III. Neural Network Training and Validation

A. Simulated Dataset

The training of the NN is performed offline on PC in MATLAB environment by means of QAT, a method to accelerate and compress the network, in which both weights and biases are quantized, during the learning phase, with a resolution of 5 bits and limited to a ±0.5 range. Before inference operations, the weights are loaded in the ASIC by means of a serial peripheral interface (SPI) and stored on local write-only SRAM cells. The NN has 1762 weights, and 8.81 kb of SRAMs are required to implement them.

The training dataset has been obtained with a Monte Carlo simulation using ANTS2, a simulation package developed for gamma camera type detectors [74]. The simulation is based on a PET detector composed by a 51 mm × 51 mm × 10 mm LYSO scintillator readout by an 8 × 8 array of 6.2 mm × 6.2 mm NUV-HD SiPMs [75]. The scintillator is wrapped with Teflon tape and coupled to the SiPM array with optical grease. The simulation emulates the interaction of 511 keV photons within the crystal, the generated scintillation light and the SiPM response. Before the training phase, a preprocessing step is carried out in which an energy filter selects only events generated by gamma photons whose total deposited energy fell within ± 5% of the 511-keV photopeak. The dataset was formed by 100 000 events collected from flood field irradiation with known position of interaction, to avoid the potential risk of overfitting that could happen if the dataset was formed by a grid of discrete positions [33]. 75% of the data has been used for training, 15% for test, and 10% for validation.

Another validation dataset has been acquired on a smaller 11×11 grid collected from pencil-beam irradiation, with 600 events/position, to assess network performances on data different from that used for training.

1) Results: The quantized network showed good positioning performances in terms of spatial resolution for the validation grid predictions, achieving an average FWHM in the x and y directions (derived from the 2-D point-spread function, 2-D PSF) of 1.22 and 1.21 mm, respectively, and a mean absolute error (MAE) of 1.66 mm.

2) Discussion: These results are comparable to the ones obtained with a not-quantized network (1.23 and 1.14 mm FWHM in the x and y directions, 1.72-mm MAE), demonstrating that 5-bit weights discretization is a good compromise between circuit complexity, occupied area, power consumption, and network performances. Also other metrics have been evaluated such as the full-width-at-tenth-maximum (FWTM), the $r_{50}\%$ and $r_{90}\%$ values, whose values are the 50% and 90% percentiles derived from the normalized cumulative distribution functions (CDFs) of the x, y and total errors. All results for the simulated dataset are summarized in Table I, comparing the NN without weight quantization (No quant.) and the NN with weight quantization (QAT).

B. Experimental Dataset

The second dataset was made of experimental data collected with a state-of-the-art monolithic PET detector consisting of a 32 mm × 32 mm × 22 mm LYSO:Ce scintillator readout by an array of 8 × 8 digital photon counters (DPCs) with pixel size of 3.2 mm × 3.87 mm [9]. The four lateral faces of the scintillator are covered with a specular reflector foil (Vikuiti ESR, 3M), whereas the top face is covered with Teflon tape. The crystal is coupled to the DPC array using a removable transparent silicon gel (Sylgard 527, Dow Corning). The measurements were acquired by [9] irradiating the crystal with a perpendicular pencil beam of 511 keV photons on a regular grid made of 128 × 128 points, for a total of 100 events per position. The first preprocessing step was to select the events for which the total energy deposited fell within the FWTM of the 511-keV photopeak. Next, a correction to obtain a homogeneous response from the DPC pixels was applied by means of LUTs build for uniformity correction from perpendicular pencil-beam and fan-beam datasets obtained during detector calibration.

In the original paper [9], this dataset was used only for validation purposes, since the calibration was performed using a different dataset obtained with a fan-beam irradiation. In this article, this set of measurement was divided into three subsets (50% training events, 10% validation events, and 40% test events). Borghi et al. [9] employed the k-nearest neighbor (k-NN) classification method for x and y position estimation, achieving an average spatial resolution in the x and y directions of about 1.7-mm FWHM/1.6-mm MAE and a total spatial resolution of 2.48-mm MAE.

1) Results: The quantized NN achieved a spatial resolution of 2.77-mm FWHM/1.55-mm MAE and 2.92-mm FWHM/1.57-mm MAE in the x and y directions, respectively, and a total spatial resolution of 2.46-mm MAE. The results obtained with a not quantized NN are 2.78-mm FWHM/1.55-mm MAE (average in the x and y directions) and a total spatial resolution of 2.44-mm MAE, thus demonstrating that the QAT did not affect the NN positioning performances. All the metrics are summarized in Table II, reporting results for the quantized NN, the not quantized NN and the k-NN method.

TABLE I
POSITIONING PERFORMANCE OF THE NN WITHOUT WEIGHT QUANTIZATION (NO QUANT.) AND WITH WEIGHT QUANTIZATION (QAT) ON THE SIMULATED DATASET

	x		y		Total	
Resolution [mm]	No quant.	QAT	No quant.	QAT	No quant.	QAT
FWHM-2D PSF	1.23	1.22	1.14	1.21	-	-
FWTM-2D PSF	3.0	2.90	2.78	3.05	-	-
$r_{50\%}$	0.61	0.63	0.61	0.64	1.06	1.13
$r_{90\%}$	2.60	2.40	2.60	2.40	3.86	3.47
MAE	1.1	1.07	1.08	1.05	1.72	1.66

TABLE II
POSITIONING PERFORMANCE OF THE NN WITH WEIGHT QUANTIZATION (NN QAT), THE NN WITHOUT QUANTIZATION (NN NO QUANT.), AND THE *k*-NN METHOD ON THE EXPERIMENTAL DATASET

	x			y			Total		
Resolution [mm]	NN QAT	NN no quant.	*k*-NN	NN QAT	NN no quant.	*k*-NN	NN QAT	NN no quant.	*k*-NN
FWHM-2D PSF	2.77	2.79	1.68	2.94	2.78	1.70	-	-	-
FWTM-2D PSF	5.9	6.02	4.76	6.12	6.11	5.02	-	-	-
$r_{50\%}$	1.05	1.04	0.82	1.08	1.06	0.84	1.86	1.84	1.62
$r_{90\%}$	3.37	3.32	3.53	3.40	3.41	3.60	4.85	4.83	5.11
MAE	1.55	1.55	1.53	1.57	1.56	1.58	2.46	2.44	2.48

2) Discussion: Better values of FWHM are achieved with the *k*-NN method; this is due to the fact that the 2-D PSFs obtained with this method do not have a Gaussian shape but have a narrow, sharp peak in the center, while the ones obtained with the NN have a Gaussian distribution. Therefore, since the two methods provide error distributions with different shapes, the spatial resolutions given by the FWHM/FWTM cannot be considered a meaningful parameter to compare their positioning performance. A more meaningful comparison can be obtained considering the $r_{50}\%$, $r_{90}\%$, and MAE values. The $r_{50}\%$ values are slightly better for the *k*-NN method, whereas the $r_{90}\%$ are slightly better for the NN method. The MAE values instead are practically equivalent for the two methods. These values show that the two methods have slightly different performance for the events which are positioned with better or worse precision (events in the $r_{50}\%$ and $r_{90}\%$ values, respectively), however they can be considered to have equivalent performance for practical applications, as shown by the MAE values. Therefore, position estimation using NNs provides a promising performance also in this case, even if the crystal used to test these methods is thick (22 mm), and therefore it is not the most suitable one for position estimation. In fact, in thick scintillators, the light distribution of events interacting in the same *x*, *y* coordinates can significantly change with the DOI.

IV. CIRCUIT DESIGN

A. Full Circuit Simulation

A preliminary simulation of the full circuit in ideal conditions implementing the NN at 10-MHz clock frequency was carried out in Cadence environment [76] for a given input event with position of interaction (–10 mm, 5 mm) within the crystal of the simulated detector. The input pattern (8 × 8 SiPM voltage signals) and the two output voltages representing the predicted *x*, *y* coordinates are reported in Fig. 6. The reconstruction error resulted equal to 0.8 mm (Euclidean distance between real and predicted coordinates), in line with the theoretical performance calculated with MATLAB.

B. Impact of Nonidealities

Despite the simplicity of the architecture, several nonideal behaviors of the electronic components, affecting the final performance, have been identified and analyzed during the design of the circuit. Particular care has been taken toward switches since they introduce charge injection, clock feedthrough and parasitic capacitance. Another criticality is the changing value of the weighting capacitance $C_{i,j}$ connected to the integrator virtual ground which affects the stability and the offset of the integrator. All these problems have been addressed by designing all components in order to minimize these unwanted effects, and characterized with post layout simulations. Parasitic capacitance values have been extracted and introduced in the NN training phase, to better match the physical circuit. A detailed description of how these problems have been addressed is outside the scope of this article, and will be object of another paper more specific on the ASIC design.

The effect of noise was carefully analyzed and minimized by design. From transient noise simulations, we have quantified the noise at the output of integrator due to thermal kTC noise of the switches frozen on capacitors and due to integrator electronic noise. Taking into account the contribution of all inputs connected to the same integrator (maximum 64 for the first network layer), the combination of both noise sources, considering 64 inputs, was estimated to be at maximum 4.7-mV rms at integrator output. To study the impact of noise on NN positioning performance, we performed a MATLAB simulation in which 10 000 inference steps were performed for the same input pattern, adding at each neuron of each layer a noise with standard deviation ranging from 0 to 5 mV rms, as shown in Fig. 7. The positioning performance is evaluated as the standard deviation of the predicted coordinates, and in the plots of Fig. 7 it is shown, for both *x* and *y*, how it varies as a function of the noise. A maximum noise of 5-mV rms produces a fluctuation at the network output of ~0.1-mm standard deviation for both coordinates, thus ~0.235-mm FWHM, i.e., below 15% if we consider a target spatial resolution better

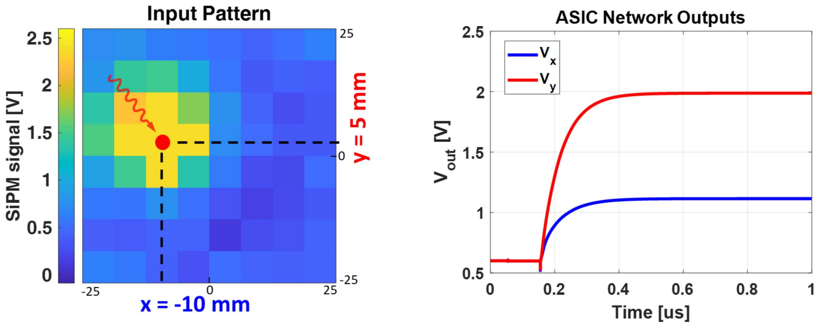

Fig. 6. Input pattern (8 × 8 SiPM voltage signals) and the two output voltages representing the *x*, *y* predicted coordinates obtained from a full circuit simulation in Cadence with 10-MHz clock frequency.

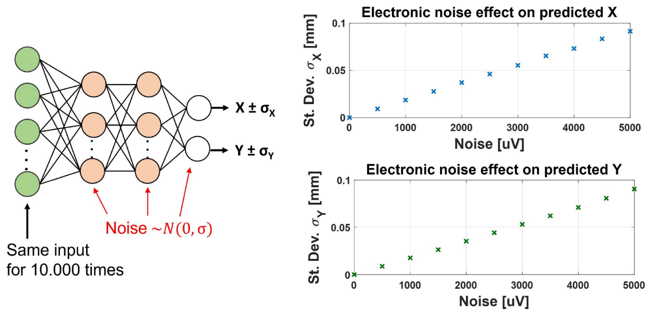

Fig. 7. MATLAB simulation to study the effect of kTC noise and integrator electronic noise on NN positioning performances: the same input pattern is given to the network 10 000 times, and at each inference step, a noise with standard deviation ranging from 0 to 5 mV is added at each neuron of each layer. The performance is evaluated as the standard deviation of the predicted coordinates.

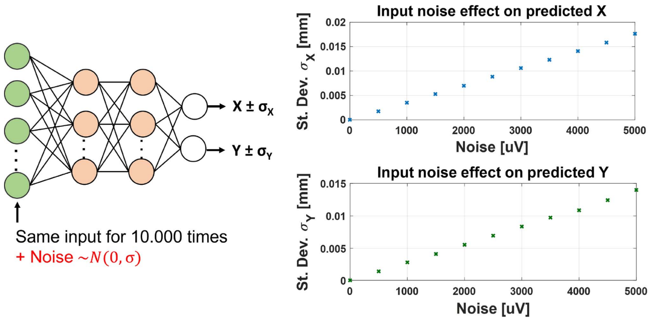

Fig. 8. MATLAB simulation to study the effect of noisy input on NN positioning performances: the same input pattern is given to the network 10 000 times, and at each inference step, a noise with standard deviation ranging from 0 to 5 mV is added at the inputs of the first layer. The performance is evaluated as the standard deviation of the predicted coordinates.

than 1.5-mm FWHM. However, since the error due to noise is statistically independent from that obtained without noise, their contributions can be summed quadratically, therefore, we can consider negligible the fluctuation produced by noise.

A similar MATLAB simulation has been performed to evaluate the impact of noisy input signals on NN reconstruction performance. The same input pattern was given to the network 10 000 times, and at each inference, a random noise from a normal distribution with zero mean and standard deviation ranging from 0 to 5 mV rms was added at the inputs of the first layer, as shown in Fig. 8. A standard deviation of the reconstructed coordinates less than 0.02 mm (∼0.05-mm FWHM) is obtained for the maximum input noise of 5-mV rms, thus negligible with respect to the target resolution.

C. Estimated Power Consumption

The ASIC is designed on Cadence environment in a 0.35-μm CMOS process node and requires an area of 23 mm^2, of which ∼7.5 mm^2 are occupied by capacitors. The chosen LSB capacitance C is of 100 fF. The choice of this technology is justified by the fact that it is compatible with most analog front-end circuit (demonstrating the potential for monolithic integration), has a lower cost compared to more recent processes, and the ASIC is a first proof-of-concept prototype. The use of a more scaled technology is envisaged in future prototypes, allowing for smaller weighting capacitors, reduced parasitic capacitances, lower latency and better efficiency.

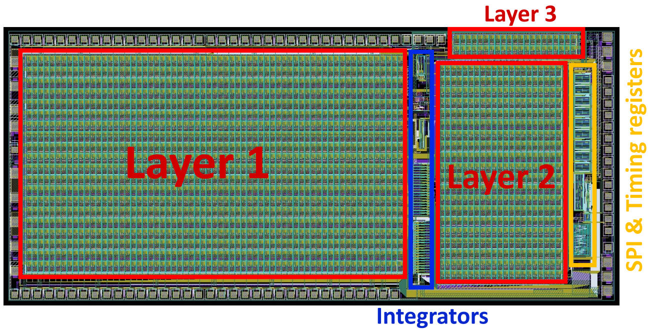

Fig. 9. Layout of the full chip implementing the analog NN in a 0.35-μm CMOS technology process node. The NN layers (crossbar arrays of switched capacitors) are highlighted with red boxes, charge integrators are in the blue box, and SPI and timing registers are in the yellow box. The total area is of 23 mm^2.

A picture of the full chip layout is shown in Fig. 9. The fundamental blocks are highlighted: the first hidden layer (with 65 × 20 neurons), the second hidden layer (with 21 × 20 neurons), and the output layer (with 21 × 2 neurons) are marked with red boxes, all the charge integrators are in the blue box, the SPI and timing registers for switches control and synchronization are in the yellow box.

If the circuit is driven with a clock frequency of 10 MHz, the minimum input-to-output processing time is estimated to be 4.6 μs, leading to a maximum 200 kHz events frequency. This corresponds to a maximum equivalent computational performance of ∼775.2 MOP/s. The estimated energy consumption to perform an inference is 38 nJ and the estimated power consumption is 7.6 mW, considering the contribution of all components involved in the MAC operation (i.e., capacitors, integrators, switches, digital timing registers for switches control and synchronization). The expected energy efficiency is 93.55 GOP/s/W, which is aligned with FPGA devices and low power GPUs designed on recent process nodes according to this recent survey [77]. If in the calculation we consider the contribution of analog elements only (i.e., without considering the timing registers), the energy efficiency results equal to 135.7 GOP/s/W. All these values have been estimated from Cadence by means of post layout simulations.

V. Discussion

The proposed analog NN ASIC for the estimation of the 2-D position of interaction of gamma photons is composed of 2 hidden layers of 20 neurons each, achieving a theoretical spatial resolution of 1.22-mm 2-D FWHM/1.66-mm MAE on a simulated detector based on a 51 mm × 51 mm × 10 mm LYSO scintillator readout by an 8 × 8 array of 6.2 mm × 6.2 mm SiPMs. We are aware that this is a relatively small network and does not match the optimal network size for position sensitivity purposes found by other groups in recent works. For example, Freire et al. [19] employed two hidden layers of 64 nodes each for the *x*-coordinate prediction, achieving a spatial resolution of 2.6-mm FWHM/1.1-mm MAE in the *x* direction on a detector based on 1 × 8 LYSO slabs of 25.8 mm × 3.1 mm × 20 mm readout by an array of 8 × 8 SiPMs. In [33] the NN contains three hidden layers of 256 neurons each and the obtained performances are 0.5-mm 2-D FWHM/1.09-mm MAE on a 50 mm × 50 mm × 16 mm LYSO crystal coupled to 64 6 mm × 6 mm SiPMs. The NN proposed in [34] has six layers and a total of 223 neurons and achieves a spatial resolution of 0.78-mm 2-D FWHM on a 25 mm × 25 mm × 8 mm LYSO crystal readout by 64 3 mm × 3 mm SiPMs. Moreover, all these networks are employed also for the prediction of DOI.

The choice of a relatively small network was mainly dictated by the chip area and the fact that this prototype is implemented with a quite old technology (0.35-μm CMOS process). In the future we envisage to implement the network in a more scaled technology, allowing us to build a larger network, which includes also the estimation of DOI, presumably using less silicon surface.

The currently achievable event rate of 200 kHz is compatible with the typical value of the maximum count rate supported by both SPECT and PET scanners, such as INSERT [78] and Hyperion II [79], respectively. They are both MRI-compatible scanners targeting mostly preclinical applications. Interestingly, recent advances of FPGA-based high-throughput inference based on Gradient Tree Boosting have reached 2 MHz [79], promising for clinical applications, with a power dissipation of the inference operation of tens of mW, currently representing the state of the art. We aim to reach the MHz range when migrating to more scaled technologies, allowing to reduce the switching cycle duration, still at significantly lower power dissipation.

The estimated energy efficiency of the chip is of 93.5 GOP/s/W, promising for this class of applications and aligned with many AI accelerators as reported in recent surveys [72], [77], such as Google TPU v2 with ∼180 GOP/s/W [80] and MIT Eyeriss chip with ∼83.1 GOP/s/W [81], even if the comparison is not fair due to the very different purpose of the devices and their different arithmetic precision, as well as the need to take into account the relevant power dissipation of the ADC required for digital and mixed-signal engines.

In general terms, the power dissipation of stationary scanners (in the order of hundreds of Watts) is mostly due to electronics and cooling. Of course, both contributions grow with an increasing number of detection modules and channels, as will happen in the ongoing trend toward scanners with larger FoV. Furthermore, there is an interplay between the two since an increasing amount of electronics in the system increases the power consumption and, correspondingly, the total heating which requires more powerful heat removal systems. Thus, the reduction of the power consumption of the electronic chain, especially of the ADC and FPGA placed in the gantry, provides a twofold advantage: the power dissipation and footprint (area and interconnection) of the readout electronics is reduced, thus also relaxing the request for the cooling system.

In this first implementation of the ASIC, the model-based training of the NN is performed offline on a PC with standard backpropagation techniques and then weights are programmed into the ASIC. This approach does not take into account the non ideal behavior of the electronic circuit, which could be addressed by having the ASIC in the forward direction of the training loop.

VI. Conclusion

In this work, we have presented the design and development of an ASIC implementing an NN by means of crossbar arrays of programmable switched capacitors, performing vector–matrix multiplications in charge domain with 5-bit resolution weights. The ASIC is intended to be used for estimating the position of interaction of gamma photons when interacting inside unsegmented scintillator coupled to an array of SiPM for PET/SPECT detectors. The (x, y) interaction coordinates are directly provided at the output of the ASIC as voltage signals proportional to the 2-D position of interaction.

The presented analog NN has the aim to tackle the bottleneck of standard Von Neumann architectures, in which the separation between memory and computational units introduces latency and high power consumption due to the large amount of data to transmit. AI algorithms have been gaining a lot of importance for compressing information and being able to deal with a large amount of data in an efficient way, still providing excellent performances.

The limitations of current hardware have been motivating the development of alternative computing paradigms, such as IMC, in which computations are performed in the same place where data is stored, leading to significant advantages in terms of speed, power, and computation efficiency.

The proposed ASIC falls in the category of analog accelerator for IMC [82]. In the specific field of application of NN to radiation detectors, this approach is particularly suitable, considering that the intrinsic information of the detector (plus possible front-end) is in the analog form and therefore its digitization is no more necessary prior to the elaboration with the NN for the scintillation event positioning. Even though this chip is application-specific, it is a programmable analog implementation of a simple feed-forward NN with weights discretized on 5 bits, and therefore it may be of interest also for other types of applications, especially the ones where bulkiness, power consumption, and latency are of critical relevance.

The ASIC has been designed on a 0.35-μm CMOS technology. The estimated energy efficiency of the chip is of 93.5 GOP/s/W. A much higher efficiency can be expected on more advanced process nodes, especially thanks to reduced parasitic capacitance and the possibility to implement a larger network to improve the throughput (operations/s). However, the choice of an older node, consolidated for low-noise analog circuits, would demonstrate the feasibility of monolithic integration of the NN in the same ASIC of the front-end filtering the current signals of the SiPMs, such as GAMMA [83] and SITH [84].

Acknowledgment

All authors declare that they have no known conflicts of interest in terms of competing financial interests or personal relationships that could have an influence or are relevant to the work reported in this article.

References

[1] E. Enlow and S. Abbaszadeh, “State-of-the-art challenges and emerging technologies in radiation detection for nuclear medicine imaging: A review,” *Front. Phys.*, vol. 11, Apr. 2023, Art. no. 1106546.

[2] S. R. Cherry, T. Jones, J. S. Karp, J. Qi, W. W. Moses, and R. D. Badawi, “Total-body PET: Maximizing sensitivity to create new opportunities for clinical research and patient care,” *J. Nucl. Med.*, vol. 59, no. 1, pp. 3–12, 2018.

[3] S. Vandenberghe, P. Moskal, and J. S. Karp, “State of the art in total body PET,” *EJNMMI Phys.*, vol. 7, pp. 1–33, Dec. 2020.

[4] A. J. Morahan et al., “Challenges in acquiring clinical simultaneous SPECT-MRI on a PET-MRI scanner,” *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 7, no. 7, pp. 755–763, Sep. 2023.

[5] D. R. Schaart, G. Schramm, J. Nuyts, and S. Surti, “Time of flight in perspective: Instrumental and computational aspects of time resolution in positron emission tomography,” *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 5, no. 5, pp. 598–618, Sep. 2021.

[6] D. R. Schaart, “Physics and technology of time-of-flight PET detectors,” *Phys. Med. Biol.*, vol. 66, no. 9, 2021, Art. no. 09TR01.

[7] S. Vandenberghe, E. Mikhaylova, E. D’Hoe, P. Mollet, and J. S. Karp, “Recent developments in time-of-flight PET,” *EJNMMI Phys.*, vol. 3, pp. 1–30, Dec. 2016.

[8] I. D’Adda et al., “A statistical DOI estimation algorithm for a SiPM-based clinical SPECT insert,” *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 6, no. 7, pp. 771–777, Sep. 2022.

[9] G. Borghi, V. Tabacchini, and D. R. Schaart, “Towards monolithic scintillator based TOF-PET systems: Practical methods for detector calibration and operation,” *Phys. Med. Biol.*, vol. 61, no. 13, p. 4904, 2016.

[10] A. J. González et al., “Next generation of the Albira small animal PET based on high density SiPM arrays,” in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf. (NSS/MIC)*, 2015, pp. 1–4.

[11] S. Krishnamoorthy, E. Blankemeyer, P. Mollet, S. Surti, R. Van Holen, and J. S. Karp, “Performance evaluation of the MOLECUBES β-CUBE—A high spatial resolution and high sensitivity small animal PET scanner utilizing monolithic LYSO scintillation detectors,” *Phys. Med. Biol.*, vol. 63, no. 15, 2018, Art. no. 155013.

[12] L. Moliner, M. J. Rodríguez-Alvarez, J. V. Catret, A. González, V. Ilisie, and J. M. Benlloch, “NEMA performance evaluation of CareMiBrain dedicated brain PET and comparison with the whole-body and dedicated brain PET systems,” *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 15484.

[13] A. J. Gonzalez et al., “Initial results of the MINDView PET insert inside the 3T mMR,” *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 3, pp. 343–351, May 2019.

[14] A. Gonzalez-Montoro et al., “Evolution of PET detectors and event positioning algorithms using monolithic scintillation crystals,” *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 5, no. 3, pp. 282–305, May 2021.

[15] M. Galasso, A. Fabbri, C. Borrazzo, V. O. Cencelli, and R. Pani, “A theoretical model for fast evaluation of position linearity and spatial resolution in gamma cameras based on monolithic scintillators,” *IEEE Trans. Nucl. Sci.*, vol. 63, no. 3, pp. 1386–1398, Jun. 2016.

[16] P. Conde et al., “Analysis of the statistical moments of the scintillation light distribution with dSiPMs,” *IEEE Trans. Nucl. Sci.*, vol. 62, no. 5, pp. 1981–1988, Oct. 2015.

[17] X. Zhang et al., “Performance of a SiPM based semi-monolithic scintillator PET detector,” *Phys. Med. Biol.*, vol. 62, no. 19, pp. 7889–7904, 2017.

[18] F. Mueller, S. Naunheim, Y. Kuhl, D. Schug, T. Solf, and V. Schulz, “A semi-monolithic detector providing intrinsic DOI-encoding and sub-200 ps CRT TOF-capabilities for clinical PET applications,” *Med. Phys.*, vol. 49, no. 12, pp. 7469–7488, 2022.

[19] M. Freire et al., “Position estimation using neural networks in semi-monolithic PET detectors,” *Phys. Med. Biol.*, vol. 67, no. 24, 2022, Art. no. 245011.

[20] “Imas project.” Accessed: Feb. 2024. [Online]. Available: https://www.fullbodyinsight.com/imas-project

[21] M. Streun, H. Nöldgen, G. Kemmerling, and S. Van Waasen, “Position reconstruction in monolithic block detectors,” in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf. Rec. (NSS/MIC)*, 2012, pp. 3212–3215.

[22] S. España, R. Marcinkowski, V. Keereman, S. Vandenberghe, and R. Van Holen, “DigiPET: Sub-millimeter spatial resolution small-animal PET imaging using thin monolithic scintillators,” *Phys. Med. Biol.*, vol. 59, no. 13, p. 3405, 2014.

[23] R. Marcinkowski, P. Mollet, R. Van Holen, and S. Vandenberghe, “Sub-millimetre DOI detector based on monolithic LYSO and digital SiPM for a dedicated small-animal PET system,” *Phys. Med. Biol.*, vol. 61, no. 5, p. 2196, 2016.

[24] M. Stockhoff, R. Van Holen, and S. Vandenberghe, “Optical simulation study on the spatial resolution of a thick monolithic PET detector,” *Phys. Med. Biol.*, vol. 64, no. 19, 2019, Art. no. 195003.

[25] K. Gong, E. Berg, S. R. Cherry, and J. Qi, “Machine learning in PET: From photon detection to quantitative image reconstruction,” *Proc. IEEE*, vol. 108, no. 1, pp. 51–68, Jan. 2020.

[26] B. Pedretti et al., “Experimental assessment of PCA and DT classification for streamlined position reconstruction in anger cameras,” *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 7, no. 4, pp. 315–325, Apr. 2023.

[27] F. Müller, D. Schug, P. Hallen, J. Grahe, and V. Schulz, “Gradient tree boosting-based positioning method for monolithic scintillator crystals in positron emission tomography,” *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 2, no. 5, pp. 411–421, Sep. 2018.

[28] F. Muller, D. Schug, P. Hallen, J. Grahe, and V. Schulz, “A novel DOI positioning algorithm for monolithic scintillator crystals in PET based on gradient tree boosting,” *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 4, pp. 465–474, Jul. 2019.

[29] P. Bruyndonckx, S. Leonard, J. Liu, S. Tavernier, P. Szupryczynski, and A. Fedorov, “Study of spatial resolution and depth of interaction of APD-based PET detector modules using light sharing schemes,” *IEEE Trans. Nucl. Sci.*, vol. 50, no. 5, pp. 1415–1419, Oct. 2003.

[30] P. Bruyndonckx et al., “Neural network-based position estimators for PET detectors using monolithic LSO blocks,” *IEEE Trans. Nucl. Sci.*, vol. 51, no. 5, pp. 2520–2525, Oct. 2004.

[31] L. Tao, X. Li, L. R. Furenlid, and C. S. Levin, “Deep learning based methods for gamma ray interaction location estimation in monolithic scintillation crystal detectors,” *Phys. Med. Biol.*, vol. 65, no. 11, 2020, Art. no. 115007.

[32] A. Sanaat and H. Zaidi, “Depth of interaction estimation in a preclinical PET scanner equipped with monolithic crystals coupled to SiPMs using a deep neural network,” *Appl. Sci.*, vol. 10, no. 14, p. 4753, 2020.

[33] M. Decuyper, M. Stockhoff, S. Vandenberghe, and R. Van Holen, “Artificial neural networks for positioning of gamma interactions in monolithic PET detectors,” *Phys. Med. Biol.*, vol. 66, no. 7, 2021, Art. no. 075001.

[34] P. Carra et al., “A neural network-based algorithm for simultaneous event positioning and timestamping in monolithic scintillators,” *Phys. Med. Biol.*, vol. 67, no. 13, 2022, Art. no. 135001.

[35] B. A. Spencer et al., “Performance evaluation of the uEXPLORER total-body PET/CT scanner based on NEMA NU 2-2018 with additional tests to characterize PET scanners with a long axial field of view,” *J. Nucl. Med.*, vol. 62, no. 6, pp. 861–870, 2021.

[36] G. Benjamin et al., "Software-based real-time acquisition and processing of PET detector raw data," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 2, pp. 316-327, Feb. 2016.

[37] F. Mateo et al., "High-precision position estimation in PET using artificial neural networks," *Nucl. Instrum. Methods Phys. Res. Sect. A, Accel., Spectrom., Detect. Assoc. Equip.*, vol. 604, nos. 1–2, pp. 366–369, 2009.

[38] Y. Wang, W. Zhu, X. Cheng, and D. Li, "3D position estimation using an artificial neural network for a continuous scintillator PET detector," *Phys. Med. Biol.*, vol. 58, no. 5, p. 1375, 2013.

[39] P. Conde et al., "Determination of the interaction position of gamma photons in monolithic scintillators using neural network fitting," *IEEE Trans. Nucl. Sci.*, vol. 63, no. 1, pp. 30–36, Feb. 2016.

[40] A. LaBella, P. Vaska, W. Zhao, and A. H. Goldan, "Convolutional neural network for crystal identification and gamma ray localization in PET," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 4, no. 4, pp. 461–469, Jul. 2020.

[41] E. Wang et al., "Deep neural network approximation for custom hardware: Where we've been, where we're going," *ACM Comput. Surv.*, vol. 52, no. 2, pp. 1–39, 2019.

[42] Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, "Recent advances in convolutional neural network acceleration," *Neurocomputing*, vol. 323, pp. 37–51, Jan. 2019.

[43] G. Akkad, A. Mansour, and E. Inaty, "Embedded deep learning accelerators: A survey on recent advances," *IEEE Trans. Artif. Intell.*, early access, Sep. 5, 2023, doi: 10.1109/TAI.2023.3311776.

[44] L. Kljucaric and A. D. George, "Deep learning inferencing with high-performance hardware accelerators," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 4, pp. 1–25, 2023.

[45] N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," in *Proc. 44th Annu. Int. Symp. Comput. Archit.*, 2017, pp. 1–12.

[46] F. Akopyan et al., "TrueNorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, Oct. 2015.

[47] "Mythic @ hot chips 2018." Accessed: Feb. 2024. [Online]. Available: https://medium.com/mythic-ai/mythic-hot-chips-2018-637dfb9e38b7

[48] "A mythic approach to deep learning inference." Accessed: Feb. 2024. [Online]. Available: https://www.nextplatform.com/2018/08/23/a-mythic-approach-to-deep-learning-inference/

[49] "Drilling into Microsoft's Brainwave soft deep learning chip." Accessed: Feb. 2024. [Online]. Available: https://www.nextplatform.com/2017/08/24/drilling-microsofts-brainwave-soft-deep-leaning-chip/

[50] "Intel FPGA architecture focuses on deep learning inference." Accessed: Feb. 2024. [Online]. Available: https://www.nextplatform.com/2018/07/31/intel-fpga-architecture-focuses-on-deep-learning-inference/

[51] S. M. Khan and A. Mann. "Ai chips: What they are and why they matter." 2020. [Online]. Available: https://cset.georgetown.edu/wp-content/uploads/AI-Chips

[52] S. Wei, X. Lin, F. Tu, Y. Wang, L. Liu, and S. Yin, "Reconfigurability, why it matters in AI tasks processing: A survey of reconfigurable AI chips," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 70, no. 3, pp. 1228–1241, Mar. 2023.

[53] D. V. Christensen et al., "2022 roadmap on neuromorphic computing and engineering," *Neuromorphic Comput. Eng.*, vol. 2, no. 2, 2022, Art. no. 022501.

[54] H. Tsai, S. Ambrogio, P. Narayanan, R. M. Shelby, and G. W. Burr, "Recent progress in analog memory-based accelerators for deep learning," *J. Phys. D, Appl. Phys.*, vol. 51, no. 28, 2018, Art. no. 283001.

[55] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnol.*, vol. 15, no. 7, pp. 529–544, 2020.

[56] X. Huang, C. Liu, Y.-G. Jiang, and P. Zhou, "In-memory computing to break the memory wall," *Chin. Phys. B*, vol. 29, no. 7, 2020, Art. no. 078504.

[57] C. Silvano et al., "A survey on deep learning hardware accelerators for heterogeneous HPC platforms," 2023, *arXiv:2306.15552*.

[58] R. Lippmann, "An introduction to computing with neural nets," *IEEE Assp Mag.*, vol. 4, no. 2, pp. 4–22, Apr. 1987.

[59] H. P. Graf et al., "VLSI implementation of a neural network memory with several hundreds of neurons," *AIP Conf. Proc.*, vol. 151, no. 1, pp. 182–187, 1986.

[60] A. J. Agranat, C. F. Neugebauer, and A. Yariv, "A CCD based neural network integrated circuit with 64K analog programmable synapses," in *Proc. IJCNN Int. Joint Conf. Neural Netw.*, 1990, pp. 551–555.

[61] Y. Tsividis and D. Anastassiou, "Switched-capacitor neural networks," *Electron. Lett.*, vol. 23, no. 18, pp. 958–959, 1987.

[62] T. Morishita, Y. Tamura, T. Otsuki, and G. Kano, "A BiCMOS analog neural network with dynamically updated weights," *IEICE Trans. Electron.*, vol. 75, no. 3, pp. 297–302, 1992.

[63] M. Holler, S. Tam, H. Castro, and R. Benson, "An electrically trainable artificial neural network (ETANN) with 10240'floating gate'synapses," in *Proc. Int. Joint Conf. Neural Netw.*, 1989, pp. 191–196.

[64] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *Neurocomputing*, vol. 74, nos. 1–3, pp. 239–255, 2010.

[65] A. Shawahna, S. M. Sait, and A. El-Maleh, "FPGA-based accelerators of deep learning networks for learning and classification: A review," *IEEE Access*, vol. 7, pp. 7823–7859, 2019.

[66] R. Machupalli, M. Hossain, and M. Mandal, "Review of ASIC accelerators for deep neural network," *Microprocess. Microsyst.*, vol. 89, Mar. 2022, Art. no. 104441.

[67] Y. E. Wang, G.-Y. Wei, and D. Brooks, "Benchmarking TPU, GPU, and CPU platforms for deep learning," 2019, *arXiv:1907.10701*.

[68] K. Seshadri, B. Akin, J. Laudon, R. Narayanaswami, and A. Yazdanbakhsh, "An evaluation of edge TPU accelerators for convolutional neural networks," in *Proc. IEEE Int. Symp. Workload Charact. (IISWC)*, 2022, pp. 79–91.

[69] T. P. Xiao, C. H. Bennett, B. Feinberg, S. Agarwal, and M. J. Marinella, "Analog architectures for neural network acceleration based on non-volatile memory," *Appl. Phys. Rev.*, vol. 7, no. 3, 2020, Art. no. 031301.

[70] P. Mannocci et al., "In-memory computing with emerging memory devices: Status and outlook," *APL Mach. Learn.*, vol. 1, no. 1, 2023, Art. no. 010902.

[71] Y. Chen, Y. Xie, L. Song, F. Chen, and T. Tang, "A survey of accelerator architectures for deep neural networks," *Engineering*, vol. 6, no. 3, pp. 264–274, 2020.

[72] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey of machine learning accelerators," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, 2020, pp. 1–12.

[73] M. S. Akhoon et al., "High performance accelerators for deep neural networks: A review," *Expert Syst.*, vol. 39, no. 1, 2022, Art. no. e12831.

[74] A. Morozov, V. Solovov, R. Martins, F. Neves, V. Domingos, and V. Chepel, "ANTS2 package: Simulation and experimental data processing for anger camera type detectors," *J. Instrum.*, vol. 11, no. 4, 2016, Art. no. P04022.

[75] A. Gola et al., "NUV-sensitive silicon photomultiplier technologies developed at Fondazione Bruno Kessler," *Sensors*, vol. 19, no. 2, p. 308, 2019.

[76] "Cadence." Accessed: Feb. 2024. [Online]. Available: https://www.cadence.com/en_US/home.html

[77] K. Guo et al. "Neural network accelerator comparison." 2022. [Online]. Available: https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/

[78] M. Carminati et al., "Clinical SiPM-based MRI-compatible SPECT: Preliminary characterization," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 4, no. 3, pp. 371–377, May 2020.

[79] K. Krueger, F. Mueller, P. Gebhardt, B. Weissler, D. Schug, and V. Schulz, "High-throughput FPGA-based inference of gradient tree boosting models for position estimation in PET detectors," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 7, no. 3, pp. 253–262, Mar. 2023.

[80] P. Teich. "Tearing apart Google's TPU 3.0 AI coprocessor." Accessed: Feb. 2024. [Online]. Available: https://www.nextplatform.com/2018/05/10/tearing-apart-googles-tpu-3-0-ai-coprocessor/

[81] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.

[82] M. L. Gallo et al., "A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference," 2022, *arXiv:2212.02872*.

[83] L. Buonanno, D. Di Vita, M. Carminati, and C. Fiorini, "Gamma: A 16-channel spectroscopic ASIC for SiPMs readout with 84-dB dynamic range," *IEEE Trans. Nucl. Sci.*, vol. 68, no. 10, pp. 2559–2572, Oct. 2021.

[84] F. Canclini, I. D'Adda, L. Buonanno, M. Carminati, and C. Fiorini, "Design of readout electronics for dose monitoring detectors in hadrontherapy," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf. (NSS/MIC)*, 2021, pp. 1–3.