IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

Received 3 September 2019; revised 13 October 2019, 28 November 2019, and 11 December 2019; accepted 13 December 2019.
Date of publication 17 December 2019; date of current version 16 January 2020.

Digital Object Identifier 10.1109/JXCDC.2019.2960307

Energy-Efficient Convolutional Neural Network Based on Cellular Neural Network Using Beyond-CMOS Technologies

CHENYUN PAN[1] (Member, IEEE), QIUWEN LOU[2], MICHAEL NIEMIER[2] (Senior Member, IEEE), SHARON HU[2] (Fellow, IEEE), and AZAD NAEEMI[3] (Senior Member, IEEE)

[1]Department of Electrical Engineering, The University of Texas at Arlington, Arlington, TX 76010 USA
[2]Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA
[3]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA

CORRESPONDING AUTHOR: C. PAN (chenyun.pan@uta.edu).

This work was supported by the Semiconductor Research Corporation (SRC) Nanoelectronics Research Initiative (NRI) Theme 2624.001 and ASCENT, one of the six centers in JUMP, a Semiconductor Research Corporation (SRC) Program sponsored by DARPA.

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

ABSTRACT In this article, we perform a uniform benchmarking for the convolutional neural network (CoNN) based on the cellular neural network (CeNN) using a variety of beyond-CMOS technologies. Representative charge-based and spintronic device technologies are implemented to enable energy-efficient CeNN related computations. To alleviate the delay and energy overheads of the fully connected layer, a hybrid spintronic CeNN-based CoNN system is proposed. It is shown that low-power FETs and spintronic devices are promising candidates to implement energy-efficient CoNNs based on CeNNs. Specifically, more than 10× improvement in energy-delay product (EDP) is demonstrated for the systems using spin diffusion-based devices and tunneling FETs compared to their conventional CMOS counterparts.

INDEX TERMS Beyond-CMOS technology, cellular neural network (CeNN), convolutional neural network (CoNN), spintronics, tunnel FETs (TFETs).

I. INTRODUCTION

CMOS technology scaling faces major challenges as we approach sub-10-nm technology nodes [1]. Steep-threshold devices, such as tunneling FETs and negative capacitance FETs (NCFETs), as well as low-voltage spintronic devices have been proposed to augment or even replace the CMOS technology. Despite major research efforts on these beyond-CMOS devices, limited performance/energy improvements have been projected compared to their CMOS counterparts for Boolean computation [2], [3]. Because many emerging beyond-CMOS devices have fundamentally different physics and offer unique characteristics, it is critically important to find novel and nontraditional circuit applications to realize their full potential.

Neuromorphic circuits have become an attractive alternative non-Boolean computing platforms in recent years, and they are shown to effectively utilize beyond-CMOS charge-based and spintronic device technologies [4]–[9]. Computing circuits that are biologically inspired has been shown to be highly efficient for tackling many complex tasks, especially in the areas of image and video processing. Massive parallel low-power computing blocks can be taken advantage of to enable energy-efficient computation [10]–[14]. Many existing proposals have investigated the neuromorphic computing systems based on beyond-CMOS devices. These systems are expected to provide significantly lower energy per operation compared to the conventional CMOS technology [9], [12]–[15].

Recently, multiple efforts have investigated the cellular neural network (CeNN) in the context of emerging information processing technologies. As representative examples, CeNNs based on graphene devices, spintronic devices, and tunnel FETs (TFETs) [16]–[22] have been studied. A variety of applications, such as noise filtering, associative memory, pattern recognition, tactile sensing, and image processing application engines have been considered [17]–[20]. Promising results were shown for these applications using emerging technologies compared to their CMOS counterparts.

A CeNN is an analog array processor architecture [23], [24] that can improve the power and performance of various computation-intensive information processing applications [25]. The underlying mathematics of a CeNN was proposed by Chua and Yang [23], and the dynamic state equation of each cellular neuron cell circuit is written as follows:

$$C_f \frac{dx_{ij}}{dt} = -\frac{1}{R_f} x_{ij} + \sum_{kl \in S_{ij}} A_{ij,kl} y_{kl} + \sum_{kl \in S_{ij}} B_{ij,kl} u_{kl} + I_{ij} \quad (1)$$

where x_{ij} is the cell state voltage, R_f is the linear feedback resistance, C_f is the linear feedback capacitance, y_{kl} are the output of the neighboring cells, and u_{kl} are the inputs of the neighboring cells. The output of each cell is defined as $y_{kl} = f(x_{kl})$, where $f(x)$ is the sigmoidal function. A_{kl} and B_{kl} are the kernels of each cell. The kernel values are equal to the synapse weights connecting two nearby cells, and I_{ij} is the input cell bias current. CeNNs are attractive as: 1) each cell is connected to only its neighbors, making the interconnections between cells local and 2) the cells and their synaptic interconnections are usually space-invariant, which makes CeNNs very suitable for CMOS very large scale integration (VLSI) implementation [26]–[29]. CeNN has shown great potential for convolutional neural network (CoNN) type of computations [30], with 8.7× and 4.3× energy-delay product (EDP) improvements compared with the state-of-the-art deep neural network (DNN) acceleration system in MNIST and CIFAR-10 data set, respectively. The CeNN architecture is well suited for the convolution operation, which is the most expensive operation in a typical CoNN. The reasons are: 1) the architecture can perform the convolution operation within one feature map all in parallel, which greatly improves the processing time; 2) the CeNN architecture supports analog computations well, which can be used to reduce energy when operating in a limited precision; and 3) the architecture is easy to realize in VLSI.

In this article, we adopt a recent architecture-level work to implement a novel way of employing cellular operations to perform the convolution, which constitutes the core computations in CoNNs [31]. This creates a unique opportunity to develop a low energy/delay mixed-signal system composed of CeNNs for realizing widely adopted CoNNs. The CeNN-based CoNN benchmarking framework presented here is generic and applicable to a wide range of beyond-CMOS charge- and spin-based devices. Comparisons among promising device choices provide insights into how to better utilize emerging technologies for energy-efficient high-performance machine learning applications. In addition, for spintronic-based systems, a hybrid CoNN design scheme is proposed to alleviate the delay and energy overhead associated with the fully connected layer in a CoNN. MNIST data set [32] is used for the benchmarking effort, but more complex data sets, such as CIFAR-10, can also be incorporated by building a wider and deeper CeNN-based CoNN.

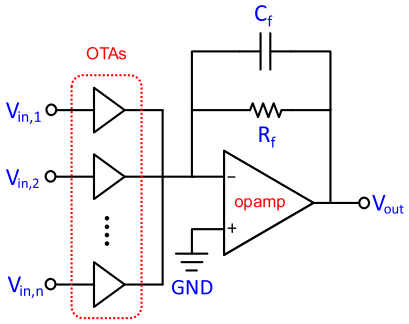

FIGURE 1. CeNN cell implemented with analog circuits.

The main contributions of this article are listed as follows.

1) We adapt a CeNN-based CoNN architecture and further improve the energy and delay per operation by using binary output states.
2) We perform a uniform benchmarking study for a wide range of charge- and spin-based devices for the first time based on the CoNN application.
3) We propose a hybrid system that uses spintronic devices for the convolution layer and charge-based devices for the fully connected layer to better utilize the advantage of spintronic devices.

The organization of this article is listed as follows. Section II illustrates two types of CeNN cell designs for charge- and spin-based devices. Section III describes the CeNN-based CoNN architecture implementation, including convolution, activation, pooling, and fully connected layers. Section IV demonstrates the functionality of the proposed CeNN-based CoNN for spintronic devices as well as the benchmarking results for both charge- and spin-based technologies. In addition, a hybrid design using CMOS and spintronic devices is also proposed and benchmarked for the spintronic implementation to alleviate the delay and energy overhead associated with the fully connected layer. Conclusions are summarized in Section V.

II. CeNN CELL DESIGN USING EMERGING TECHNOLOGIES

In this section, we describe two types of CeNN cell designs that are implemented with a variety of beyond-CMOS devices, including charge-based FETs and spintronic devices. In [44], the digital CeNN implementation is shown to be more energy- and time-consuming due to the fact that each convolution operation requires reading data, multiple stages of summation, and storing data in the registers. Therefore, in this work, the same CeNN implementation for CoNN is briefly discussed in the Supplementary material.

A. CHARGE-BASED CeNN CELL IMPLEMENTATION

For the charge-based implementation, we use analog circuits illustrated in Fig. 1 to achieve the CeNN dynamics in (1). The first-order differential equation is implemented by using an analog opamp-based integrator. The opamp design follows a standard differential-input single-ended output

TABLE 1. Charge-based device characteristics.

Device Name Abbreviation	Device Full Name	Vdd (V)	Ion (μA/μm)	Ioff (μA/μm)	Cg (aF) @ W = 60nm	Subthreshold Slope (mV/Dec)	Bias Current (μA/μm)
CMOS HP [33]	CMOS High Performance FET	0.73	1805	0.101	61.39	70.6	13.5
CMOS LV [34]	CMOS Low Voltage FET	0.3	53.0	0.003	40.27	70.1	0.4
HomJTFET [35]	Homojunction Tunneling FET	0.2	24.5	0.063	49.11	77.3	1.2
HetJTFET [36]	Heterojunction Tunneling FET	0.4	416.7	0.005	49.11	61.3	1.5
ThinTFET [37]	2D Heterojunction Interlayer Tunnel FET	0.2	858.0	0.002	59.63	50.0	158
GaNTFET [38]	Gallium Nitride tunneling FET	0.3	88.6	0.002	37.96	65.4	0.4
TMDTFET [39]	Transition Metal Dichalcogenide Tunnel FET	0.2	76.0	0.002	59.03	103	15.0
FEFET [40]	Ferroelectric FET	0.73	1293	0.103	54.61	73.2	11.5
NCFET [41]	Negative Capacitance FET	0.3	410.8	0.002	68.63	64.0	7.5
MITFET [39]	Metal Insulator Transition FET	0.5	1531	0.301	137.6	81.0	21.5
GpnJ-Vg2 [42]	Graphene pn Junction FET	0.3	164.2	15.00	1573	289	49.6
GpnJ-Vg3 [42]	Graphene pn Junction FET	0.3	535.7	3.261	1573	135	41.8
vdWFET-BP [43]	Van der Walls FET	0.3	34.3	0.003	41.80	73.3	0.3

seven-transistor operational amplifiers [45]. The synapse weights, namely kernels A and B in (1), are achieved by using operational transconductance amplifiers (OTAs) [46]. We used OTAs with quantized tail transistor widths to achieve digitally programmable synapses. Memory cells, such as SRAM, are used to control the gates of the tail transistors to achieve a variety of functionalities.

The bias of the amplifier is assumed to be V_{out}/R_f to ensure desired output swing, where the feedback resistance, R_f, in the CeNN needs to be set properly to achieve the correct functionality of the CoNN. For a general CoNN operation, the desired output, y, is written as follows:

$$y = \sum w \cdot u + i \tag{2}$$

where u is the input from the previous layer, w the kernel weight, and i the input bias. To map CoNN operation to the CeNN dynamic in (1), the kernel B can be written as $w \cdot g_m$, where g_m is the transconductance of the OTA, and $I = i \cdot g_m$. At a steady state, the cell voltage x is expected to reach the desired output y, and the gradient of the cell voltage should reach 0, which gives

$$0 = -\frac{1}{R_f} y + \sum w \cdot g_m \cdot u + i \cdot g_m. \tag{3}$$

With (2) and (3), the feedback resistance, R_f, should be set as $1/g_m$ so that the cell voltage x settles to y. The feedback capacitance, C_f, is the summation of the input capacitances of nearby OTAs, which is written as follows:

$$C_f = C_{\text{OTA}} \cdot N_{\text{b}}(2N_s + 1) \tag{4}$$

where C_{OTA} is the output capacitance of the OTA, N_s is the weight kernel size of 9, and N_b is the number of bits of the synapse, which is assumed to be 4 in this work.

In this article, 13 representative charge-based FET devices are investigated. The characteristics of emerging devices as well as two baseline CMOS devices, namely, CMOS HP and CMOS LV devices, are listed in Table 1. Many of the beyond-CMOS charge-based FETs have an advantage in terms of the steep subthreshold thanks to their distinct operation principles, such as the negative capacitance effect in NCFET [41], Klein tunneling effect in graphene p–n junction FET (GpnJ) [52], and band-to-band tunneling effect in TFET. Note that the variability effects for analog circuits, such as the threshold voltage variation, have not been considered. The main purpose of this article is to give an upper bound on the potential of each emerging technology and investigate its energy and delay per operation. This helps us to identify promising device candidates so that a more thorough investigation can be done for those devices in the future.

B. SPINTRONIC CeNN CELL IMPLEMENTATION

For the spintronic implementation, we use magnetic components complemented with CMOS peripherals to create a complete driving and sensing circuitry. Three types of spintronic CeNN cell designs are adopted from the previous work and illustrated in Fig. 2, including spin Hall effect devices, spin diffusion devices, and domain wall devices [44].

The dynamic switching behavior of magnets in the neuron is captured by solving the Landau–Lifshitz–Gilbert (LLG) equation with an added spin-transfer-torque term [53], [54]. The magnet magnetization, $\vec{m}$, can be written as follows:

$$\frac{d\vec{m}}{dt} = -\gamma\mu_0[\vec{m} \times \vec{H}_{\text{eff}}] + \alpha\left[\vec{m} \times \frac{d\vec{m}}{dt}\right] + \frac{\vec{I}_{S,\perp}}{eN_s} \tag{5}$$

where μ_0 is the permittivity, $\vec{H}_{\text{eff}}$ is the effective magnetic field, γ is the gyromagnetic ratio, α is the damping factor of the magnet, e is the elementary charge, N_s is the number of magnetons, and $\vec{I}_{S,\perp}$ is the perpendicular spin-polarized current. This spin-polarized current is expressed as $i_0 \cdot (\sum_{kl \in S_{ij}} A_{ij,kl} y_{kl} + \sum_{kl \in S_{ij}} B_{ij,kl} u_{kl} + I_{ij})$, where i_0 is the unit spin-polarized current when the kernel value in A or B is unity. The output and input voltages of nearby CeNN cells and the weights of synapses connecting those cells determine the amplitude and the direction of the spin-polarized current.

The inputs of the CeNN cell, V_{in} and $\bar{V}_{\text{in}}$, are received from inputs, u_{kl}, of nearby CeNN cells. These inputs are connected to n-type driving transistors with various driving

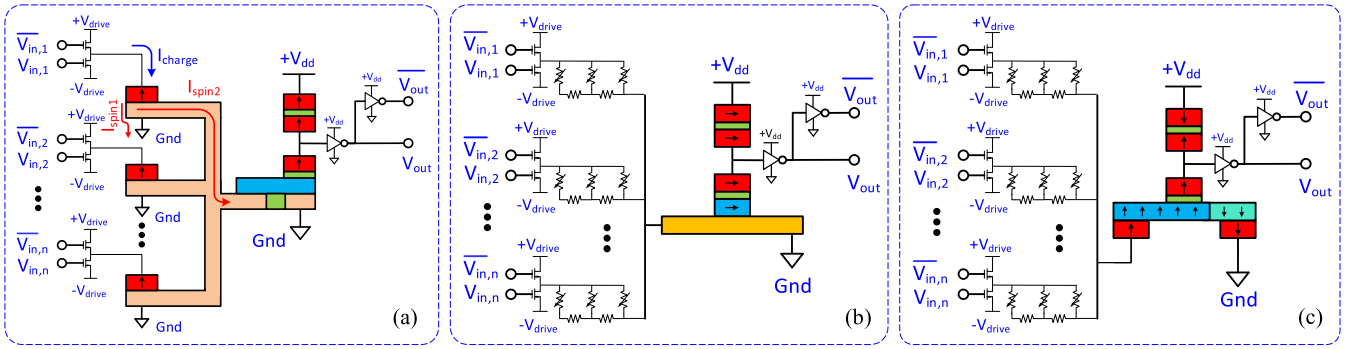

FIGURE 2. CeNN cell implemented with magnetic synapses and neurons based on (a) spin diffusion, (b) spin Hall effect, and (c) domain wall motion as the writing mechanisms.

TABLE 2. Spintronic device and cell characteristics.

Device Name Abbreviation	Device Full Name	Driving Voltage (mV)	Critical Switching Current (μA)	CeNN Cell Delay (ns)
SD [47]	Spin Diffusion Device	100	39	21.9
SD-HA [48]	Spin Diffusion Device with Heusler Alloy	100	17.8	1.9
SHE [49]	Spin Hall Effect Device	100	38.7	4
SHE-CC [50]	Spin Hall Effect Device with Copper Collector	80	38.7	2
SHE-YIG [50]	Spin Hall Effect Device with Yttrium Iron Garnet	35	25.8	1
DW [51]	Domain Wall Motion Device	50	0.8	6.8

sizes that are determined by the absolute values of kernels, B_{ij}. The input voltage turns on one of the driving MOSFET, which generates a charge current. For the spin diffusion-based CeNN cell design shown in Fig. 2(a), the charge current flowing through the fixed magnet gets spin-polarized. The injected spins diffuse toward the free magnet in the neuron. Both the direction of the magnetization the magnet and the charge current determines the polarity of the spin current. For the spin Hall effect-based CeNN cell design shown in Fig. 2(b), the charge current flowing through the heavy metal insert spin-orbit torque to the cell magnet and determine the magnetization direction. For the domain wall-based CeNN cell design shown in Fig. 2(c), the charge current drives the domain wall and determines the MTJ state. The readout circuitry for three spintronic implementations is identical and designed by using two MTJs [17]. If the bottom MTJ is at parallel (or antiparallel) configurations, the voltage at the input of the inverter becomes low (or high) accordingly. Then, the complementary voltages are generated with two inverters to drive the synapses in nearby CeNN cells. The magnet dynamics in each cell is simulated numerically by solving the LLG equation shown in (5). In this article, six representative spintronic devices with different materials are investigated. The driving voltage and switching characteristics of different devices are listed in Table 2.

III. BENCHMARKING FRAMEWORK: CeNN-BASED CoNN CIRCUIT AND ARCHITECTURE

In this section, we describe the benchmarking framework by introducing a CeNN-based CoNN that can be widely used for general machine learning applications.

A. NEURAL NETWORK DESIGN

Using the CeNN cell building blocks described in Section II, we have developed a CeNN-friendly CoNN for the MNIST recognition task [55], where the system classifies each handwritten digit (0–9) that is represented by a 28 × 28 pixel image. All computational kernels are restricted to a CeNN friendly size of 3 × 3. Although larger kernels are consistent with CeNN theory (a neighborhood's radius r can be >1), large kernels are infrequently utilized to avoid connectivity challenges. Also, Szegedy *et al.* [56] suggest that smaller kernels can lead to fewer parameters/higher accuracy during training—which maps well to CeNN hardware. The resulting network is shown in Fig. 3 and consists of four channels, where different convolution kernels are employed in each channel (highlighted in blue). In this work, we do not reuse the CeNN cell across different kernels or layers. Each convolution operation has a dedicated CeNN cell. For more complex tasks and data sets, one can store the weight in the memory and load the weight each time before reusing the CeNN cells. To train the synapse weights of the neural network, we used the Caffe toolsets [57].

B. LAYER IMPLEMENTATIONS IN CeNN-BASED CoNN

We show in this section that the core CoNN computation can be readily mapped to a CeNN hardware. In the previous work, the analog implementation of a CeNN-based CoNN has been demonstrated by using CMOS devices to perform energy-efficient non-Boolean computation [30], where the bit precision of the synapse weights is 4. In this section, we describe a few major changes to the activation and pooling layers and apply it to both charge-based analog and spintronic implementation.

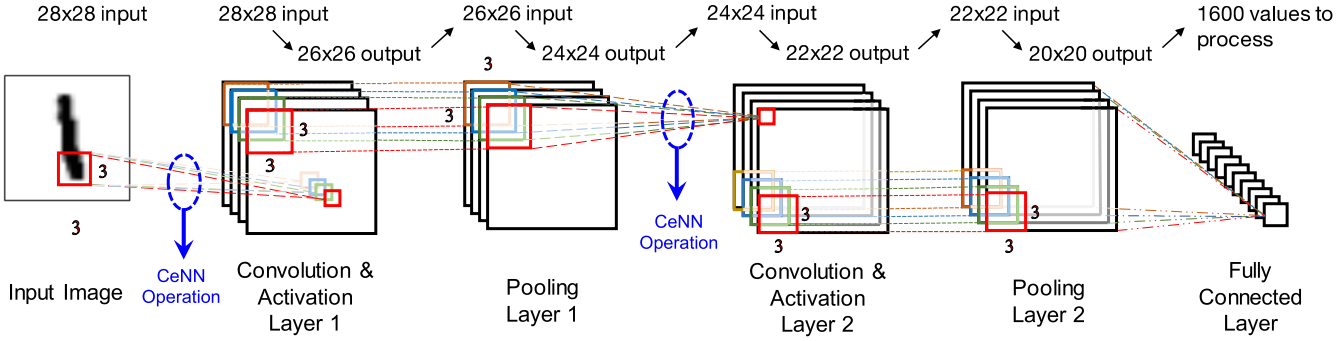

FIGURE 3. **CeNN-based CoNN architecture for the MNIST application.**

1) CONVOLUTIONAL LAYER

Convolutional layers are used to detect and extract features on the input data by convolving the input with convolutional kernels. One CeNN can implement a convolutional operation in a straightforward manner by simply setting the values of the B-kernel to be the values of the convolutional kernel.

2) ACTIVATION LAYER

In the previous work, the activation layer used a rectified linear unit (ReLU) operation because it is the most commonly used nonlinearity activation function in deep learning applications. For the CeNN-based CoNN, the ReLU operation requires two linear operations that first shift all values down by one followed by shifting up by one. This method is effective but induces two extra steps. In addition, for the spintronic implementation, because the CeNN cell design illustrated in Fig. 2 has binary output features due to the bistability of magnets, the ReLU activation function cannot be achieved by using the aforementioned two-step CeNN operations. Therefore, in this work, we modified the activation function from ReLU to a sigmoidal-like function with a steep transition to match the output characteristic of spintronic devices. For a fair comparison, the analog implementation uses comparators to achieve the same activation function.

3) POOLING LAYER

Pooling operations are employed in-between successive layers to reduce the spatial information so that the network parameters are reduced. Here, we discuss the implementation of a widely used pooling function in CoNN—max pooling. We compare the value of the center pixel with all its neighbors. If the center pixel is larger, we keep the center pixel; if the center pixel is smaller, we replace the center pixel with its neighbor pixel. In the previous work, we developed a series of CeNN kernels to realize the max pooling operation, which requires 16 steps. Due to the fact that we binarize the output after the activation layer, the max-pooling function can be achieved by simply using or operations.

4) FULLY CONNECTED LAYER

The operation of the fully connected layer can be defined as a pixelwise dot product between a weight matrix and a feature map. The result can be used as a classification result. For weight matrix sizes larger than 3×3 (which is typical in CoNN implementation), the fully connected layer cannot be efficiently implemented by CeNN. To overcome this challenge, one can leverage digital circuits, such as adders and multipliers, to perform the fully connected layer function. In this article, we have adopted the arithmetic logic unit (ALU) from previous work to perform Boolean addition and multiplication for the fully connected layer [2].

IV. SIMULATION AND BENCHMARKING RESULTS

In this section, a variety of beyond-CMOS devices are benchmarked using the same simulation framework described in Section III. For magnetic devices, we perform detailed numerical LLG simulations to demonstrate the functionality of the spintronic CeNN-based CoNN for the MNIST digit recognition problem. One of the key nonideal effects, thermal noise, is taken into account in the simulation. For the charge-based devices, due to limited inputs from different research centers, we did not investigate the nonideality of each type of device and its impact on the overall application-level accuracy. The main purpose of this article is to give an upbound of the potential of each emerging technology and investigate its energy and delay per operation. This helps us to identify promising device candidates so that a more thorough investigation can be done for those devices in the future. In the end, a hybrid scheme is proposed to better utilize the potential of spintronic-based implementation.

A. FUNCTIONALITY DEMONSTRATION

Based on the CeNN cell design and neural network design presented in Sections II and III, we perform the benchmarking for a generic CoNN. In this article, we perform a benchmarking for the MNIST digit recognition problem as a case study. The same benchmarking framework is applicable to all DNN architectures for various applications.

As we discussed in Section III, the spintronic activation layer uses the sigmoidal-like function to take into account the binary switching characteristics of the magnets. The activation function can be expressed as $1/(1 + e^{-\alpha x})$. Based on the modified CoNN, retraining is performed to obtain the updated weights using Caffe [57]. We use a relatively large α of 20 to capture the binary characteristics of the magnets used

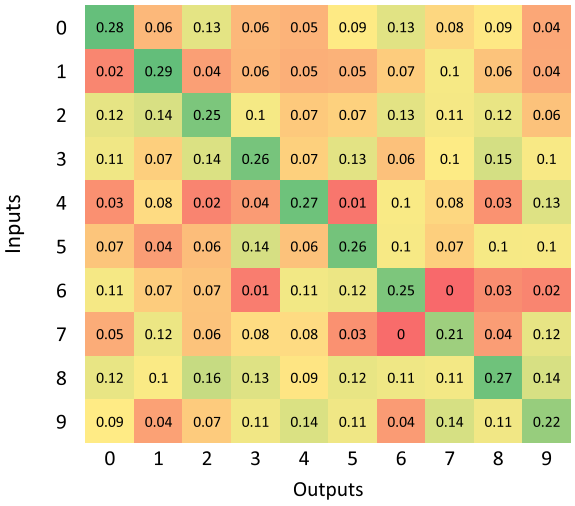

FIGURE 4. **Functional demonstration of spintronic CeNN-based CoNN system for the MNIST application.**

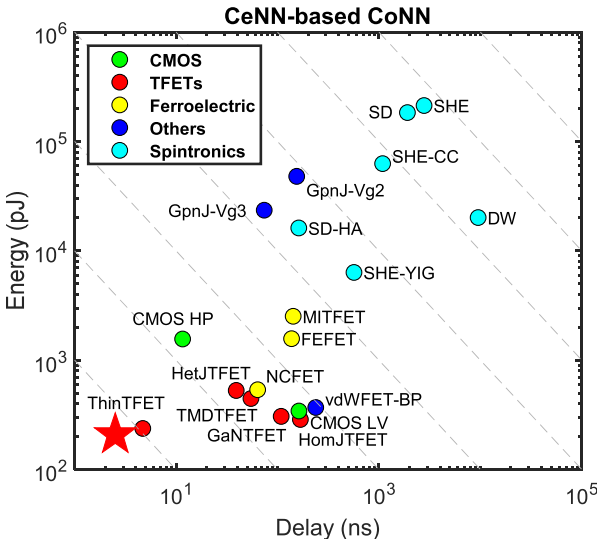

FIGURE 5. **Comparison of energy and delay per MNIST recognition task among various beyond-CMOS technologies with analog and spintronic implementations. Green, red, yellow, and blue dots represent charge-based devices, whose expansion of each device is listed in Table 1. Cyan points represent spintronic devices, whose expansion of each device is listed in Table 2. The details of each device flavor are described in the previous benchmarking article [2]. The red star at the bottom left shows the preferred corner.**

in spintronic implementations. After the retraining, the recognition accuracy of the modified network reaches 97%. Then, the corresponding weights are exported to MATLAB, and detailed numerical LLG simulations are performed to precisely capture the magnet dynamics in each CeNN cell under the influence of the thermal noise. Due to the time-consuming numerical simulations, 100 samples are taken to demonstrate the functionality of the spintronic CeNN-based CoNN. The average output values after softmax for ten digits versus the input pattern is shown in Fig. 4, where strong correlations between the normalized output values and input labels are observed. The overall recognition accuracy also reaches 97%.

B. PERFORMANCE BENCHMARKING

In this section, based on the validated weight information, we perform a thorough benchmarking of key performance metrics of the systems implemented by a variety of beyond-CMOS charge- and spintronic devices. The modeling of energy and delay per CeNN operation under given weights are adopted from [44]. For spintronic devices, different writing mechanisms are investigated, such as spin diffusion, spin Hall effect and domain wall motion, corresponding to the schematics shown in Fig. 2.

Fig. 5 shows the overall energy dissipation and delay per image recognition task for five categories of beyond-CMOS devices. In general, tunneling FETs, shown as red dots, dissipate less energy compared to other charge- and spin-based devices. The main reason is the steep subthreshold slopes of TFETs, which allows ultralow supply voltages as shown in Table 1. In addition, a steep subthreshold slope leads to a larger g_m at a small bias current, which reduces the required feedback resistance, R_f, based on (3). A small R_f helps us to reduce the delay constant and improve the settling time of the cellular operation. Among five TFETs being investigated in this article, the system implemented by ThinTFET stands out and achieves the best EDP due to its large driving current under a small supply voltage. The low energy systems technology (LEAST) center provides the current–voltage characteristics, which are obtained through atomistic simulations [2].

In Fig. 5, one can also observe that the systems using spintronic devices (light blue points) are located at the top right corner, which is away from the preferred corner. This is mainly because the energy and delay associated with the last fully connected layer are prohibitively large. As described in Section III, due to the limited number of inputs in each CeNN cell, we use digital Boolean logic circuits with adders and multipliers to perform multiplications and additions in the fully connected layer, assuming the number of output bits is 12. Because of the relatively slow response time of a magnet compared to the delay of charging or discharging the gate of an FET, spintronic devices are very inefficient in performing Boolean logic functions.

To better visualize the computation cost associated with the fully connected and convolutional layers, we broke down the delay and energy dissipation of the two parts for various device options as shown in Fig. 6. The six spintronic implementations are located on the top right corner of Fig. 6(a), indicating the large computation cost of the Boolean operations. For the convolution layer shown in Fig. 6(b), spintronic systems are more competitive, and over one order of magnitude improvement can be observed for spin diffusion-based devices using Heusler alloy (SD-HA) magnets. This is because: 1) a single magnet can accomplish the functionality of a neuron, and the convolution operation can be achieved in a fast and efficient way through a single cellular operation and 2) the spintronic device operates at a low supply

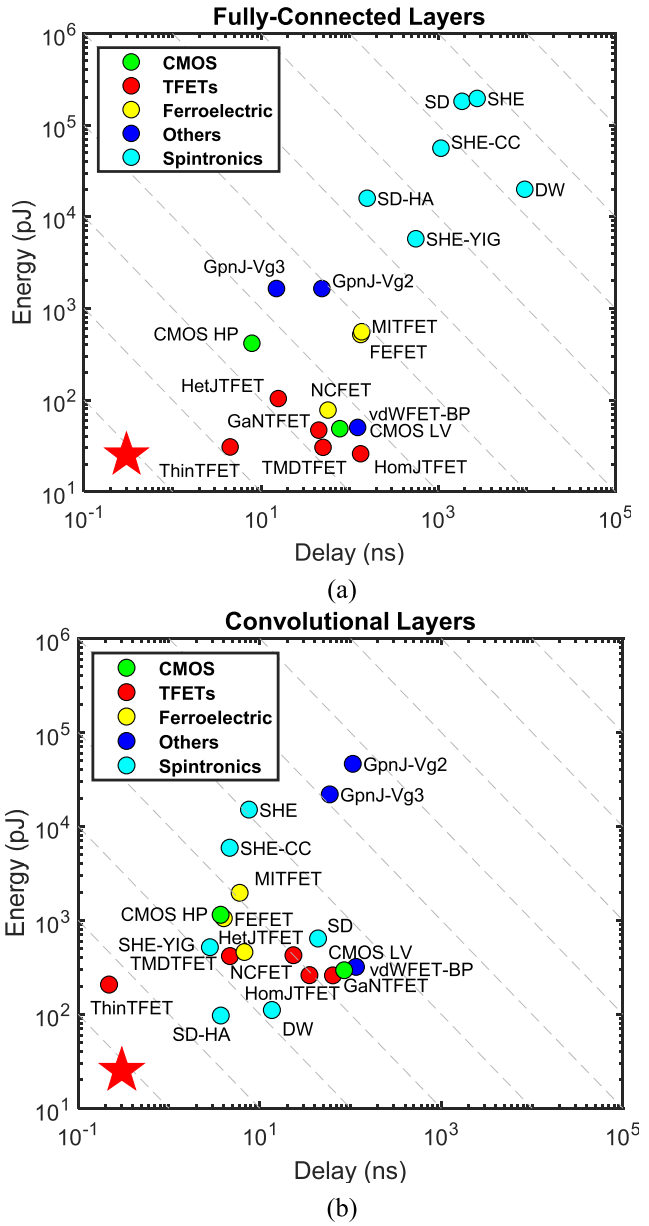

FIGURE 6. Energy versus delay for (a) fully connected layer and (b) convolution layers in a CeNN-based CoNN system for a variety of beyond-CMOS technologies.

voltage, which further improves the energy efficiency of the computation.

For the system implemented by charge-based FETs, the energy dissipation per operation associated with the convolutional layers is higher compared to the energy for fully connected layer due to a large number of convolution operations involved. For the delay, similar to the spintronic systems, the delay of charge-based systems is limited by the fully connected layer. Here, we assume there are 100 multipliers available for the Boolean computation. If we allocate more resources for the multiplier, the delay associated with the fully connected layers can be further reduced, but this will come with a larger footprint area overhead.

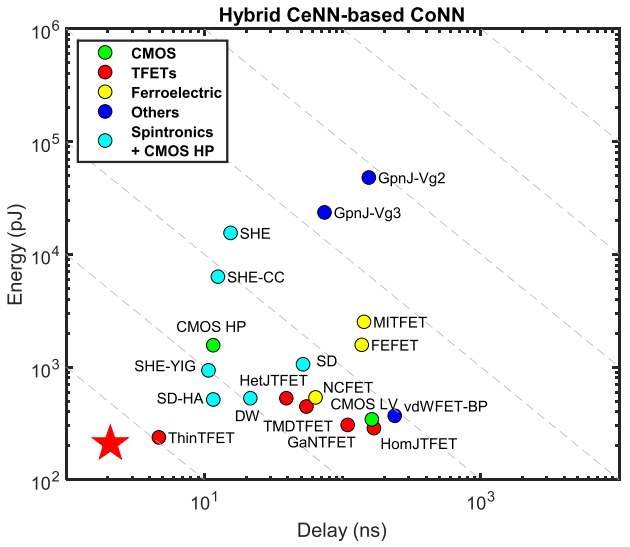

FIGURE 7. Energy versus delay of CeNN-based CoNN performing MNIST digit recognition task for a variety of beyond-CMOS technologies. Here, cyan dots represent the hybrid design using a combination of CMOS HP devices for implementing the fully connected layer and spintronic devices for the convolution layers in the CoNN.

For charge-based FETs with an extra ferroelectric switching time (yellow dots), such as FEFET, MITFET, and NCFET, their relative positions in the convolutional layer plot [Fig. 6(a)] shift largely toward the preferred corner compared to the ones shown in the fully connected layer plot [Fig. 6(b)]. This is mainly because of the extra ferroelectric polarization switching time. This extra delay can be much larger than the intrinsic switching delays of FETs during the Boolean logic computation; however, for the CeNN-based CoNN application, the product of the feedback resistance and capacitance, $R_f C_f$, dominates the settling time, which overshadows the extra ferroelectric switching delay.

C. HYBRID CONFIGURATION FOR SPINTRONIC CeNN-BASED CoNN IMPLEMENTATION

To further improve the performance of spintronic systems (cyan points in Fig. 5), we propose to implement a hybrid configuration. The proposed hybrid system uses CMOS HP devices to perform the Boolean computation in the fully connected layer, and for the rest of the layers, the same spintronic CeNN cells are used. Because each spintronic CeNN cell implementation has voltage output characteristics as described in Section II, the output of the pooling layer can be directly connected with the fully connected layer using charge-based CMOS devices.

To quantify the potential improvement offered by the proposed hybrid system, the updated energy versus delay comparison for various emerging technologies is shown in Fig. 7. One can observe that compared with the data in Fig. 5, the hybrid systems using spintronic and CMOS HP devices show promising results. Their positions shift significantly toward the bottom left preferred corner. Multiple hybrid spintronic systems outperform their counterparts using

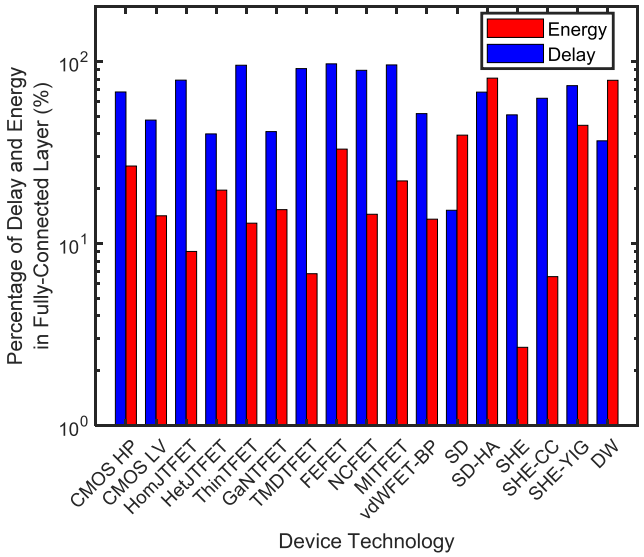

FIGURE 8. Percentage of delay and energy dissipation associated with the fully connected layer for a variety of beyond-CMOS technologies.

conventional CMOS HP devices. For the system using SD-HA material, it provides a 4× EDP reduction because of the low critical current requirement of the magnetic material. With the continuous improvement of spintronic devices and the discovery of new materials, the performance of spintronic circuits is expected to be further improved.

To identify the limitation and bottleneck of the proposed hybrid CeNN-based CoNN system, Fig. 8 shows the percentage of delay and energy associated with the fully connected layer. One can observe that the delay and energy cost of the Boolean computation in the hybrid spintronic-CMOS systems decreases significantly compared to the results shown in Fig. 6. However, for ultralow-power systems, such as the ones using SD-HA and SHE-YIG, the delay and energy are still dominated by CMOS HP devices in the fully connected layer. The main reason SHE device has a much smaller percentage of energy dissipation associated with the fully connected layer is because this device consumes a lot of energy during the convolution operation, as can be seen from Fig. 6(b). Therefore, the overall energy dissipation is more dominated by the convolution layer instead of the fully connected layer. One possible option to alleviate the overhead is to use low-power charge-based devices, such as TFETs, to replace the CMOS HP devices in the hybrid systems, which will further improve the overall EDP of the spintronic system.

V. CONCLUSION

This article presents a uniform benchmarking for non-Boolean computation for the CoNNs based on CeNNs. A variety of charge-based and spintronic beyond-CMOS device technologies are used to implement an energy-efficient CeNN cell. The functionality of the spintronic CoNN system is demonstrated through numerical simulations for the MNIST application. A hybrid spintronic CeNN-based CoNN system is proposed to alleviate the delay and energy overheads of the fully connected layer implemented by spintronic devices. It is shown that spintronic devices have the potential to efficiently implement CoNNs, where 4× improvement in EDP is projected for the system using SD-HA or ThinTFETs compared to its conventional CMOS counterpart.

REFERENCES

[1] K. J. Kuhn, "CMOS scaling for the 22 nm node and beyond: Device physics and technology," in *Proc. Int. Symp. VLSI Technol., Syst. Appl.*, Apr. 2011, pp. 1–2.

[2] C. Pan and A. Naeemi, "An expanded benchmarking of beyond–CMOS devices based on Boolean and neuromorphic representative circuits," *IEEE J. Explor. Solid-State Computat. Devices Circuits*, vol. 3, pp. 101–110, Dec. 2017.

[3] D. E. Nikonov and I. A. Young, "Benchmarking of beyond–CMOS exploratory devices for logic integrated circuits," *IEEE J. Explor. Solid-State Computat. Devices Circuits*, vol. 1, pp. 3–11, Dec. 2015.

[4] H. Markram, "The blue brain project," *Nature Rev. Neurosci.*, vol. 7, no. 2, pp. 153–160, Feb. 2006.

[5] P. A. Merolla *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

[6] A. Calimera, E. Macii, and M. Poncino, "The human brain project and neuromorphic computing," *Funct. Neurol.*, vol. 28, no. 3, p. 191, 2013.

[7] K. Moon *et al.*, "High density neuromorphic system with $Mo/Pr_{0.7}Ca_{0.3}MnO_3$ synapse and NbO_2 IMT oscillator neuron," in *IEDM Tech. Dig.*, Dec. 2015, pp. 17.6.1–17.6.4.

[8] G. W. Burr *et al.*, "Neuromorphic computing using non-volatile memory," *Adv. Phys. X*, vol. 2, no. 1, pp. 89–124, 2017.

[9] A. Sengupta and K. Roy, "Encoding neural and synaptic functionalities in electron spin: A pathway to efficient neuromorphic computing," *Appl. Phys. Rev.*, vol. 4, no. 4, Dec. 2017, Art. no. 041105.

[10] D. Monroe, "Neuromorphic computing gets ready for the (really) big time," *Commun. ACM*, vol. 57, no. 6, pp. 13–15, Jun. 2014.

[11] S. Venkataramani, A. Ranjan, K. Roy, and A. Raghunathan, "AxNN: Energy-efficient neuromorphic systems using approximate computing," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, 2014, pp. 27–32.

[12] J. Grollier, D. Querlioz, and M. D. Stiles, "Spintronic nanodevices for bioinspired computing," *Proc. IEEE*, vol. 104, no. 10, pp. 2024–2039, Oct. 2016.

[13] G. Srinivasan, A. Sengupta, and K. Roy, "Magnetic tunnel junction based long–term short–term stochastic synapse for a spiking neural network with on–chip STDP learning," *Sci. Rep.*, vol. 6, no. 1, Sep. 2016, Art. no. 29545.

[14] D. Zhang, L. Zeng, Y. Zhang, W. Zhao, and J. O. Klein, "Stochastic spintronic device based synapses and spiking neurons for neuromorphic computation," in *Proc. IEEE/ACM Int. Symp. Nanosc. Archit. (NANOARCH)*, Jul. 2016, pp. 173–178.

[15] C. D. Schuman *et al.*, "A survey of neuromorphic computing and neural networks in hardware," 2017, *arXiv:1705.06963*. [Online]. Available: https://arxiv.org/abs/1705.06963

[16] M. Darwish, V. Calayir, L. Pileggi, and J. A. Weldon, "Ultracompact graphene multigate variable resistor for neuromorphic computing," *IEEE Trans. Nanotechnol.*, vol. 15, no. 2, pp. 318–327, Mar. 2016.

[17] C. Pan and A. Naeemi, "A proposal for energy–efficient cellular neural network based on spintronic devices," *IEEE Trans. Nanotechnol.*, vol. 15, no. 5, pp. 820–827, Sep. 2016.

[18] I. Palit, X. S. Hu, J. Nahas, and M. Niemier, "TFET-based cellular neural network architectures," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, Sep. 2013, pp. 236–241.

[19] I. Palit, Q. Lou, N. Acampora, J. Nahas, M. Niemier, and X. S. Hu, "Analytically modeling power and performance of a CNN system," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2015, pp. 186–193.

[20] I. Palit, B. Sedighi, A. Horvath, X. S. Hu, J. Nahas, and M. Niemier, "Impact of steep-slope transistors on non-von Neumann architectures: CNN case study," in *Proc. Design, Automat. Test Eur. Conf. Exhib. (DATE)*, 2014, pp. 1–6.

[21] Y. Bai, S. Hu, R. F. Demara, and M. Lin, "A spin–orbit torque based cellular neural network (CNN) architecture," in *Proc. Great Lakes Symp. VLSI (GLSVLSI)*, 2017, pp. 59–64.

[22] A. W. Stephan, J. Hu, and S. J. Koester, "Performance estimate of inverse Rashba–Edelstein magnetoelectric devices for neuromorphic computing," *IEEE J. Explor. Solid-State Computat. Devices Circuits*, vol. 5, no. 1, pp. 25–33, Jun. 2019.

[23] L. O. Chua and L. Yang, "Cellular neural networks: Theory," *IEEE Trans. Circuits Syst.*, vol. CAS-35, no. 10, pp. 1257–1272, Oct. 1988.

[24] L. O. Chua and L. Yang, "Cellular neural networks: Applications," *IEEE Trans. Circuits Syst.*, vol. CAS-35, no. 10, pp. 1273–1290, Oct. 1988.

[25] L. Chua and T. Roska, "The CNN paradigm," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 40, no. 3, pp. 147–156, Mar. 1993.

[26] K. Karahaliloglu and S. Balkir, "Bio-inspired compact cell circuit for reaction-diffusion systems," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 52, no. 9, pp. 558–562, Sep. 2005.

[27] P. Kinget and M. Steyaert, "A programmable analog cellular neural network CMOS chip for high speed image processing," *IEEE J. Solid-State Circuits*, vol. 30, no. 3, pp. 235–243, Mar. 1995.

[28] J. Kowalski, "0.8 μm CMOS implementation of weighted-order statistic image filter based on cellular neural network architecture," *IEEE Trans. Neural Netw.*, vol. 14, no. 5, pp. 1366–1374, Sep. 2003.

[29] T. Roska and A. Rodriguez-Vazquez, "Toward visual microprocessors," *Proc. IEEE*, vol. 90, no. 7, pp. 1244–1257, Jul. 2002.

[30] Q. Lou, C. Pan, J. McGuinness, A. Horvath, A. Naeemi, and X. S. Hu, "A mixed signal architecture for convolutional neural networks," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 15, no. 2, pp. 19:1–19:26, Mar. 2019.

[31] A. Horvath, M. Hillmer, Q. Lou, X. S. Hu, and M. Niemier, "Cellular neural network friendly convolutional neural networks—CNNs with CNNs," in *Proc. Design, Automat. Test Eur. Conf. Exhib. (DATE)*, Mar. 2017, pp. 145–150.

[32] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the Web]," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.

[33] International Technology Roadmap for Semiconductors (ITRS). *High-Performance (HP) Logic Technology Requirements*. Accessed: 2012. [Online]. Available: http://www.itrs2.net/

[34] R. Kim, U. E. Avci, and I. A. Young, "Source/drain doping effects and performance analysis of ballistic III–V n-MOSFETs," *IEEE J. Electron Devices Soc.*, vol. 3, no. 1, pp. 37–43, Jan. 2015.

[35] A. C. Seabaugh and Q. Zhang, "Low–voltage tunnel transistors for beyond CMOS logic," *Proc. IEEE*, vol. 98, no. 12, pp. 2095–2110, Dec. 2010.

[36] H. Kroemer, "A proposed class of hetero-junction injection lasers," *Proc. IEEE*, vol. 51, no. 12, pp. 1782–1783, Dec. 1963.

[37] M. O. Li, R. Yan, D. Jena, and H. G. Xing, "Two-dimensional heterojunction interlayer tunnel FET (Thin–TFET): From theory to applications," in *IEDM Tech. Dig.*, Dec. 2016, pp. 19.2.1–19.2.4.

[38] A. Seabaugh *et al.*, "Steep subthreshold swing tunnel FETs: GaN/InN/GaN and transition metal dichalcogenide channels," in *IEDM Tech. Dig.*, Dec. 2015, pp. 35.6.1–35.6.4.

[39] H. Ilatikhameneh, Y. Tan, B. Novakovic, G. Klimeck, R. Rahman, and J. Appenzeller, "Tunnel field–effect transistors in 2-D transition metal dichalcogenide materials," *IEEE J. Explor. Solid-State Computat. Devices Circuits*, vol. 1, pp. 12–18, Dec. 2015.

[40] S. L. Miller and P. J. Mcwhorter, "Physics of the ferroelectric nonvolatile memory field effect transistor," *J. Appl. Phys.*, vol. 72, no. 12, pp. 5999–6010, Dec. 1992.

[41] S. Salahuddin and S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nano Lett.*, vol. 8, no. 2, pp. 405–410, Feb. 2008.

[42] S. Chen *et al.*, "Electron optics with p-n junctions in ballistic graphene," *Science*, vol. 353, no. 6307, pp. 1522–1525, Sep. 2016.

[43] S. S. Sylvia, K. Alam, and R. K. Lake, "Uniform benchmarking of low–voltage van der Waals FETs," *IEEE J. Explor. Solid-State Computat. Devices Circuits*, vol. 2, pp. 28–35, Dec. 2016.

[44] C. Pan and A. Naeemi, "Non–Boolean computing benchmarking for beyond–CMOS devices based on cellular neural network," *IEEE J. Explor. Solid-State Computat. Devices Circuits*, vol. 2, pp. 36–43, Dec. 2016.

[45] P. R. Gray, P. Hurst, R. G. Meyer, and S. Lewis, *Analysis and Design of Analog Integrated Circuits*. Hoboken, NJ, USA: Wiley, 2001.

[46] A. R. Trivedi and S. Mukhopadhyay, "Potential of ultralow–power cellular neural image processing with Si/Ge tunnel FET," *IEEE Trans. Nanotechnol.*, vol. 13, no. 4, pp. 627–629, Jul. 2014.

[47] B. Behin-Aein, D. Datta, S. Salahuddin, and S. Datta, "Proposal for an all-spin logic device with built-in memory," *Nature Nanotechnol.*, vol. 5, no. 4, pp. 266–270, Apr. 2010.

[48] A. K. Nayak *et al.*, "Design of compensated ferrimagnetic Heusler alloys for giant tunable exchange bias," *Nature Mater.*, vol. 14, no. 7, pp. 679–684, Jul. 2015.

[49] S. Datta, S. Salahuddin, and B. Behin-Aein, "Non-volatile spin switch for Boolean and non-Boolean logic," *Appl. Phys. Lett.*, vol. 101, no. 25, Dec. 2012, Art. no. 252411.

[50] S. Sayed, V. Q. Diep, K. Y. Camsari, and S. Datta, "Spin funneling for enhanced spin injection into ferromagnets," *Sci. Rep.*, vol. 6, no. 1, Sep. 2016, Art. no. 28868.

[51] D. Morris, D. Bromberg, J.-G. J. Zhu, and L. Pileggi, "mLogic: Ultra-low voltage non-volatile logic circuits using STT-MTJ devices," in *Proc. 49th Annu. Design Autom. Conf.*, 2012, pp. 486–491.

[52] V. V. Cheianov and V. I. Fal'ko, "Selective transmission of Dirac electrons and ballistic magnetoresistance of *n–p* junctions in graphene," *Phys. Rev. B, Condens. Matter*, vol. 74, no. 4, Jul. 2006, Art. no. 041403.

[53] D. Ralph and M. Stiles, "Spin transfer torques," *J. Magn. Magn. Mater.*, vol. 320, no. 7, pp. 1190–1216, Apr. 2008.

[54] D. Berkov and J. Miltat, "Spin-torque driven magnetization dynamics: Micromagnetic modeling," *J. Magn. Magn. Mater.*, vol. 320, no. 7, pp. 1238–1259, Apr. 2008.

[55] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015, *arXiv:1512.00567*. [Online]. Available: https://arxiv.org/abs/1512.00567

[57] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 675–678.