

Neuro-memristive Circuits for Edge Computing: A Review

Olga Krestinskaya, *Student Member, IEEE*, Alex James, *Senior Member, IEEE* and Leon O. Chua, *Fellow, IEEE*

Abstract—The volume, veracity, variability and velocity of data produced from the ever increasing network of sensors connected to Internet pose challenges for power management, scalability and sustainability of cloud computing infrastructure. Increasing the data processing capability of edge computing devices at lower power requirements can reduce several overheads for cloud computing solutions. This paper provides the review of neuromorphic CMOS-memristive architectures that can be integrated into edge computing devices. We discuss why the neuromorphic architectures are useful for edge devices and show the advantages, drawbacks and open problems in the field of neuro-memristive circuits for edge computing.

Index Terms—Memristors, Memristor circuits, Neural Networks, Cellular neural network, Convolutional neural network, Long short-term memory, Hierarchical temporal memory, Spiking neural networks, Deep neural networks

I. INTRODUCTION

THE increase in the number of edge devices such as mobile phones, and wearable electronics connected to Internet drives the scale-up of intelligent data applications. Edge computing is broadly defined as the method used for moving the control of data processing from centralized core computing nodes such as high-performance computing servers to the last edge nodes of the Internet where data is collected and connected to the physical world [1], [2]. The high velocity and volume of data generated lead to the need to scale up data centers, and puts added pressure on lowering energy consumption. However, the inability to linearly scale power with existing CMOS technology prompts us to look at neuromorphic computing architectures that can be used in edge devices and possibly useful for replacing hardware in cloud computing platforms. It is expected that in 2-5 years the edge computing technologies will be in the main stream [3], along with machine learning, Internet of Things (IoT) and smart electronics, mutually contributing to each other areas growth [4], [5].

The development of the neuro-memristive circuits that can be integrated to edge computing devices is an open research problem. Neuromorphic computing is inspired from the biological concepts of human brain processing that has the potential to replace traditional von Neumann computing paradigms. In the more than Moore's era of device scale-up and architectures, memristive circuits and computing architectures is one of the promising solutions [6]. Memristors provide

various advantages, such as scalability, small on-chip area, low power dissipation, efficiency and adaptability [7], [8].

In this paper, the correlation between neuromorphic memristive architectures with the edge computing trends is illustrated, we discuss the different set of neuromorphic architectures for the edge computing that can be integrated directly to the edge devices. We illustrate the most recent approaches to implement neuron cell and synaptic connections and show the correlation with biological concepts. We present the clear overview of various neuromorphic architectures, such as different types of neural networks [9], Hierarchical Temporal Memory (HTM) [10], [11], Long Short-Term Memory (LSTM) [12], [13], learning architectures and circuits for memory-based computing and storage. We discuss the advantages and primary challenges in the simulation and implementation of such architectures. Also, we present the main drawbacks and challenges that should be improved in existing neuromorphic architecture to use them in edge computing applications.

The paper is organized as the following. Section II provides the overview of edge computing on hardware and edge computing architectures. Section III provides the review of neuron models, relates the biological concepts to the existing neuron architectures and covers the most common circuits for hardware implementation of CMOS-memristive synapses and neuron cells. Section IV introduces various neuromorphic architectures that can be used for edge computing application. Section V illustrates the advantages, issues and open problems of the CMOS-memristive architectures. Section VI concludes the paper.

II. EDGE DEVICES AND EMERGING NEURAL COMPUTING

Figure 1 shows the overall concept of the edge computing system. The sensors in the edges of the concept map collect the data for processing in the edge devices, which in essence move part of information processing and computing tasks from cloud to edge devices. The increased demand on the edge devices to process information in intelligent and useful ways triggers the development of emerging hardware and edge AI computing.

The real-time data produced by the ever increasing number of sensors in edge devices pushes for near-sensor computing for various intelligent information processing applications. There is an emerging market of artificial intelligence chips in edge devices for utilizing machine learning and neural networks [14], [15]. The information from the sensor is converted to digital domain by analog to digital converter, followed by filtering methods and co-processors for implementing different neural network configurations [14], [15]. However, with major

O. Krestinskaya is a graduate student and A.P. James is a chair in Electrical and Computer engineering at Nazarbayev University. Email: apj@ieee.org

L.O Chua is Professor Emeritus in Electrical Engineering and Computer Sciences at University of California Berkeley.

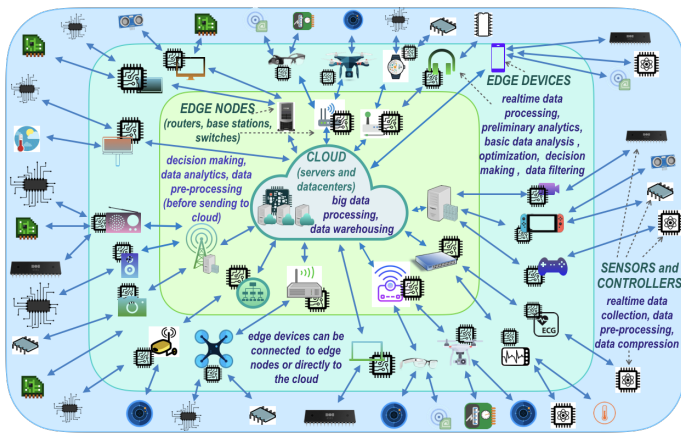


Fig. 1. Overall concept of edge computing system.

issues in scaling the devices to sub-10nm range, emerging devices such as memristors become promising to increase the speed and on-chip area. Further, these emerging devices also promote the analog domain processing of information as many neural networks in hardware can be mapped to memristive array based computing architectures [16].

Mobile devices largely has driven the growth in high-performance logic and low-power digital logic chips in the last several years. And the limitation and challenges in device scaling has forced the community to move towards neural computing solutions that can incorporate more than Moore's law [17] and beyond CMOS technologies [18], [19] as a key aspect of future hardware development. In edge devices, such as mobiles, the key computing drivers are towards having higher performance and more functionality at lower cost and energy which is constrained by battery. Several hardware technology aspects drives this development, they are: Logic technologies, Ground rule scaling, Performance boosters, Performance-power-area (PPA) scaling, 3D integration, Memory technologies, DRAM technologies, Flash technologies and Emerging non-volatile-memory (NVM) technologies such as memristors. The key performance benchmarks for node scaling for edge devices in more than Moore's era integration in the next 2-3 years includes [20]: (1) increasing the operating frequency by 15% relative to the scaled supply voltage, (2) for a given performance reduce the energy per switching by 35%, (3) reduce the area on chip footprint by 35%, and (4) reduce the scaled die cost by 20%, while limit wafer cost to increase within 30%.

Memory and logic technologies together shape the development of near sensor neural computing solutions for edge devices. Memristive devices are emerging non-volatile memories that offer several potential features to support the growth in this field. There are several non-volatile variants of memristor devices such as magnetic or MRAM [21], phase-change or PCRAM [22], and resistive or ReRAM [23] that can be used for building neural networks. The two-terminal resistive structure requires a selector device such as a transistor to program these devices in an array. The two common use of memristors in a neural computing paradigm is as a memory and as a dot-product computing unit [24]. The main purpose

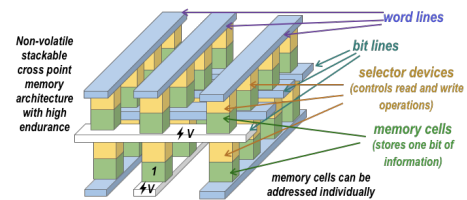


Fig. 2. 3D XP Memory Architecture [25].

of memristor as a memory is as a storage unit for weights during the learning stages in digital or discrete analog domain processing. While, a memristor crossbar array can be used for computing the dot product between the input and weights in a neural network layer in analog domain.

The development of high density crossbar memristor architecture has been limited by the lack of a good and energy efficient selector device. Being a resistive device, memristors such as ReRAM require either bi- or unipolar operation for programming to a particular state. The 3D XP memory shown in Fig. 2 [25] is been a promising direction to solve this bottleneck, and the major issue that remains is the device to device variability of the resistive state. Even with variability, the neural networks have shown robust performances, as during the learning phase, any variability in the states translates to the variability in weights, which are compensated by the learning algorithm to find the optimal set of weights that works best for the given neural network configuration.

The growth in hardware for edge computing is driven by Internet of things, where the sensors-humans-computers collaborate to provide efficient and useful intelligent application [1]. The data analysis for these application often needs to be fast, and also need to ensure security and privacy. Hardware level security is an essential advantage offered by the emerging devices [26] that can be integrated into edge devices. The progress in NVM memristor devices and arrays due to its lower operating voltages, compatibility with CMOS devices and faster speeds allows to develop a large variety of energy and area efficient neural networks configurations [16], [27].

In the edge computing concept (Fig. 1), the data processing is shifted from the data centers to the edge devices [28]. The edge computing relies on billions of various devices connected to the Internet. Each device collects the information and can process this data locally. The data processed on edge level is collected in the aggregation nodes at the intermediate fog level that incorporates the networking devices, aggregation devices, and gateways required for sending processed data to the cloud data centers [1], [2]. Cloud is on the top level of the data processing containing data warehouses, which is responsible for large data processing. The edge computing is a basis for IoT systems, which incorporates the ideas of smart devices, smart vehicles, and connected systems and can be extended to a system of systems solutions involving big data analytics. The development of IoT networks and amount of data transferred and processed in cloud stresses the limits of the data centers. If the current trends continue developing at the same pace, in few decades, the amount of energy required to process the ever growing data will overload the bandwidth requirements,

and cloud computing requirements to a point that it would not be feasible to meet the demands of speed, and cost [29].

The main idea of edge computing is the local processing of the data, which does not require sending of significant amount of data to the servers. All these decision making and processing mechanisms should be performed in low power levels. This removes the need to have complex data centers. The edge computing becomes more relevant because the power required for data processing in the data centers on servers increased significantly in the last few years. And if the growth of processed data continues, it would increase the costs for powering the data centers to support of same speed and amount of data processed on servers. As all the edge devices are limited in terms of on-chip area and low power consumption requirements, the conventional von Neumann architectures with traditional CMOS devices become less feasible for such purposes in the long-term as transistor scalability is expensive and energy per computation saturates. The neuromorphic non von Neumann architectures discussed in Section IV are considered to be a promising solution for energy-related issues and optimization of such systems. Moreover, neuromorphic architectures can be used to solve the cloud computing energy-related issues in memory and processing units and to achieve energy-efficient computing [30].

The distributed nature of the edge computing architectures allows to integrate neural chips as co-processing units within the edge devices. The neural chips make use of neuron models inspired from the biological understanding of neuronal behaviour and function. The neuron models are used to build different types of neural network configurations that can mimic functions and capacity of human brain. There are several neural architectures such as deep learning neural network (DNN) [31], [32], convolutional neural network (CNN) [33], [34], long short term memory (LSTM) [12], [13], hierarchical temporal memories (HTM) [10], [11] and generative adversarial networks (GAN) [35] that has grown prominence in the last decade.

Edge processing often involves real-time localized data processing. Therefore, the primary goal of the edge computing is to make edge devices more intelligent, faster and less power hungry. Also, it is essential to consider the issues related to communication protocols, bandwidth, and correlation of data from all edge devices. It is important to make the device more intelligent and understand which information should be processed. Therefore, the learning process [9] in neuromorphic systems is essential.

The memristive neuromorphic architectures aim to reduce the processing power, which allow integrating these architectures to the edge devices. The lower energy consumption increases the battery life, allows to pack more computing hardware modules and also decreases the overall cost of computation. The cost-effectiveness is achieved, because the memristor-based neuromorphic architectures require a smaller amount of memory for processing and networks can be learned to understand the information rather than storing and retrieving it using energy consuming hardware and software algorithms. The learning process in neuromorphic architectures allows achieving faster processing time. In memristive hardware

architectures, the learning process is slow, while the decision making and processing of data after learning is very fast. Once the memristive neuromorphic architecture is learned, the information processing on local edge devices can be performed quickly. Also, the faster data processing can be achieved using analog learning architectures, which are useful for near-sensor processing. The analog neuromorphic architectures [9] can be integrated directly to the sensors avoiding intermediate data conversion stage.

The data security issues are addressed in memristive neuromorphic architectures because the information processing is performed at a hardware level, where encryption level is high [26], [36]. There are growing incidents for hacking on chip data in digital hardware [37]. Neuromorphic architectures encode the data, and the memristive weights are learned, so it is impossible to predict the weights. Therefore, the natural encoding process is performed, and the system is more secured. The neuromorphic architectures are more robust to variations because of the learning process and weight adjustment. The interoperability can be ensured because the networks become more adaptable to the process variations in chip. Therefore, decision fusion and collaborative sensing also can be performed between the chips to ensure higher security levels. In addition, memristor based key generators can be incorporated into the chip to implement functional data security algorithms [26], [36], [38]–[40].

III. NEURON MODELS

In this section, we focus on the memristive models of neuron cells and synaptic connections that can be adapted and scaled for the edge computing applications.

A. Inspiration from biological concepts

Neuromorphic circuits and architectures attempt to mimic different types of biological neural networks responsible for information processing in human brain [41], [46]. The biological neuron architecture is shown in Fig. 3 (a). A biological neuron consists of the soma (cell body) with many dendrites that serve as connections to the other neurons and carry the information. The axon (output of the neuron) collects the information from all the dendrites and transmits it to the other neurons. The transmission of a signal from one neuron to another happens through the synapses. Synapses can either reinforce or inhibit the transmitted signals [41]. The neuron fires (generates the output response), if the information that is collected in the axon exceeds the particular threshold [42].

The equivalent structural and mathematical representation of biological neuron is shown in Fig. 3 (b) [47] and Fig. 3 (c). The neuron models can be divided into two categories: (1) simple threshold logic based linear neuron based models, where the neuron is presented as a most straightforward linear computing unit, and (2) dendritic threshold non-linear neuron based models, which has more complex computing units and is inspired by recent works [43].

The simplest threshold logic based linear neuron model is known as McCulloch-Pitts neuron model [48] and Rosenblatt's perceptron [49]. Fig. 3 (b) and Eq. 1 shows the threshold

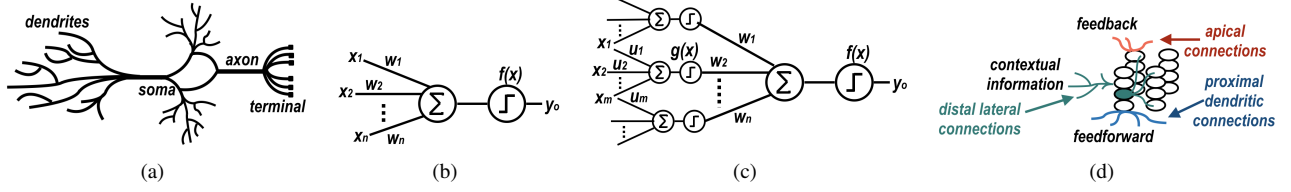


Fig. 3. (a) Biological neuron [41], (b) threshold logic based linear neuron model [41], [42], (c) dendritic threshold non-linear neuron model [43], (d) HTM neuron [44], [45].

logic based linear neuron model. The synapses are represented as weighted connections [42]. The parameter w_j represent the weights of the synapses, and y_o is a neuron output. The central concept of this model is that the weighted summation of the inputs x_j is higher than the threshold θ . This threshold determines the neuron firing.

$$y_o = f\left(\sum_{j=1}^n w_j x_j\right) \quad (1)$$

The particular case proposed in [49] is shown in Eq. 2, where the hard threshold function is used as an activation function.

$$y_o = \begin{cases} +1 & \sum_{j=1}^n w_j x_j \geq \theta \\ -1 & \text{otherwise} \end{cases} \quad (2)$$

In the dendritic threshold non-linear neuron model the dendrites of the neuron can be nonlinear. Each dendritic unit in the neuron consists of various subunits (dendritic branches), and neurons are represented as complex computing unit [43]. Fig. 3 (c) and Eq. 3 shows the structure of non-linear dendritic neuron model. A single dendrite can have multiple inputs and specific threshold function.

$$y_o = f\left(\sum_{j=1}^n w_j g_j\left(\sum_{i=1}^m u_i x_i\right)\right) \quad (3)$$

Comparing to threshold linear neuron model, like perceptron, which fails to compute particular functions, threshold non-linear neuron can compute linearly non-separable functions.

The volatility principle in human brain-inspired architectures is also important. The research work [50] claims that it is of importance not only to remember important data but also forget the unnecessary information. An HTM neuron emulates this process. HTM neuron is a particular case of dendritic threshold non-linear neuron model recently proposed to mimic functionality of pyramidal neurons [51] in human neocortex [44], [45]. The HTM neuron is shown in Fig. 3 (d). The neuron cell has three different inputs: feedforward, feedback, and contextual inputs. The feedforward input corresponds to the synapses of proximal soma known as proximal dendritic connections. The feedback inputs correspond to apical connections learned from the previous inputs, and the contextual inputs correspond to distal connections that connect different cells.

B. Memristive circuit as a synapse

1) *Single memristor as a synapse*: Most of the implementations of the neuron models propose to use memristor as a synapse. The least complex representations of the synapse in memristive architectures is a single memristor (1M) structure. The single memristor synapses in a memristive crossbar array are shown in Fig. 4 (a). The 1M structure is more efficient in terms of on-chip area and power consumption. The recent works attempt to use 1M synapses for neural networks to avoid additional CMOS elements in the architectures [8], [59], [60]. However, the neuromorphic circuits with 1M synapses usually required additional control circuits and suffered from sneak path problems. Moreover, the update process of the memristor values in such structures requires complex switch circuits, which disconnect the memristors from presynaptic and postsynaptic neurons and connect the input signals used for memristor programming. Also, such configurations do not allow to obtain negative synaptic weights, and additional circuits should be involved to obtain the negative weights in neural networks.

2) *Synapses with two memristors*: The alternative to 1M synapses is the synapses with two memristors (2M) shown in Fig. 4 (b) [52], [61], [62]. This architecture doubles the size of the crossbar and requires complex postsynaptic neurons. However, this allows implementing negative weights of the synapses. In 2M structure the weight of the synapse is represented as $W_{ij} = G_{ij}^+ - G_{ij}^-$, where G_{ij}^\pm is an effective conductance of a memristor [52], [61].

The alternative 2M synapse with PCMO memristors is shown in [63]. In this particular example, memristors are connected to long-term depression (LTD) and long-term potentiation (LTP) neurons and correspond to LTD and LTP operations, which occur during particular periods of time. When the synapse is potentiated, only the LTP memristor conductance is increase, while LTP memristor remain unchanged, and vice versa. This allows to remove the effects of asymmetric changes of the resistance level from R_{ON} to R_{OFF} and R_{OFF} to R_{ON} , avoiding abrupt changes in overall resistance of the synapse comprised of the resistances of two devices.

The 2M synapses, where two memristors are connected in series, are presented in [64]. In this work, the synapse is presented by two types of devices: diffusive memristor device $SiO_xNy : Ag$ (a device based on silver nanoparticles in a dielectric film that can be used as a selector device [64] or even neuron [65]) and drift memristor device TaO_x (usual non-volatile device). The synapse was designed to realize dynamic behavior, LTD and LTP of biological synapses.

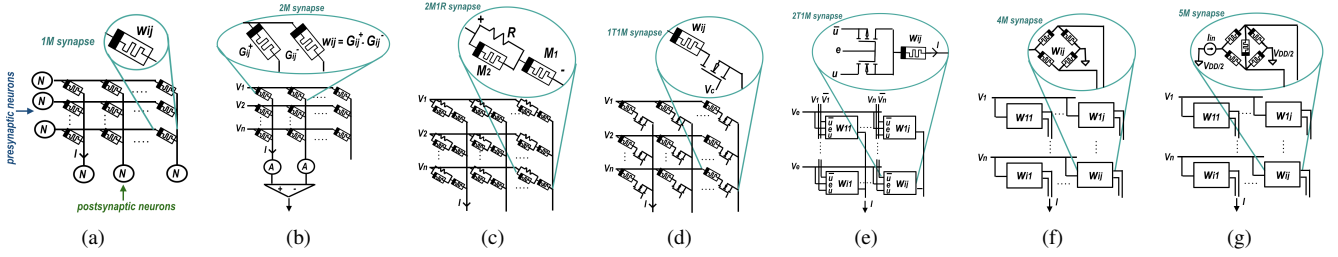


Fig. 4. Memristive synapses: (a) 1M synapse in a crossbar array, (b) 2M synapses [52], (c) 2M1R synapse [53], (d) 1T1M synapses [47], [54], (e) 2T2M synapses [55], (f) 4M synapse [56], [57] and 5M synapse [58].

The two memristor one resistor (2M1R) synapse is shown in Fig. 4 (c). The research work [53] proposes the modified dynamic synapse for SNN based on the two memristors and the resistor adjusted for TaO_x devices, which includes temporal transformations and static weight and helps to realize the spiking behavior in large-scale simulations.

3) *Synapses with transistors*: The memristive synapses with transistors are also popular because the transistor is used as a switch, especially for read and update cycles. The synapse with one transistor and one memristor (1T1M) is shown in Fig. 4 (d) [47], [54]. This architecture is one of the possible solutions for sneak path problems. The synapse with two transistors and one memristor (2T1M) is illustrated in Fig. 4 [55], [66]. While 1T1M architecture is used to control memristor switching, program the memristor within a crossbar and eliminate sneak path problems, 2T1M also allows to control the sign of the memristor, as it is connected to two inputs: original and inverted input signal. The enabling signal e controls the switching of the CMOS transistors. The transistors control the current flowing through the memristor and voltage across the memristor. The parameter e represents the enable signal. If $e = 0$, the state variable of memristor does not change. If $e = V_{DD}$ or $e = -V_{DD}$, the current is flowing either through NMOS transistor or PMOS transistor, respectively. The enable signal is used to control the direction of current and to update the memristor value. This also allows achieving negative and positive sign of memristor weight. In this circuit, it is important to ensure that the transistor is in a linear state. The drawback of such circuit is a size of the synapse, which is appropriate for small-scale problems [67], and can be a critical issue for large-scale edge computing systems.

4) *Memristor bridge synapses*: The other type of synaptic weight implementations is a bridge arrangement. The memristor-bridge synapse with 4 memristors (4M) shown in Fig. 4 (f) was tested in various neural network architectures and applications [56], [57]. The circuit consists of 4 memristors that form Wheatstone bridge-like circuit and is able to represent zero, positive, and negative synaptic weights. To increase the resistance of M_2 and M_3 and decrease of resistance of M_1 and M_4 , positive pulse should be applied as an input and vice versa. The weight is positive, if $\frac{M_2}{M_1} > \frac{M_4}{M_3}$. The negative weight can be formed as $\frac{M_2}{M_1} < \frac{M_4}{M_3}$. A zero weight is formed as $\frac{M_2}{M_1} = \frac{M_4}{M_3}$. This ensures the implementation of positive and negative weights and allows to change the weight

sign, which depends on the direction of the current.

C. Neuron cell models

1) *Integrate and fire neuron model*: The earliest neuron cell models are based on capacitors that emulate the membrane of a biological neuron and integrate current [7]. One of the basic and first neuron models is Integrate and Fire (I&F) neuron model. In this model, single membrane capacitance sums the currents flowing into the neuron from all the synapses and membrane resistance causes the leakage of the membrane current [74]. However, due to the large on-chip area and power consumption, such neurons are not applicable for large-scale circuits and edge devices, where the power consumption is limited. Even the novel I&F neuron circuits proposed recently [75]–[79] cannot be extended for the use in the large-scale systems due to the number of the components.

There are only a few attempts to use the I&F based neuron models in large-scale architectures. The modified I&F neuron used for neural network implementation is shown in Fig. 5 (a) [68]. The neuron circuit consists of current integration part with capacitor C_u , spike generation Schmitt trigger circuit, reset circuit and control circuit for current input range and injection. When the voltage is applied to the terminals of transistors M_1 and M_2 , the input current I_{in} is injected to the leaky integration part of the neuron through the current mirror. This current is integrated and leaked through M_3 . Then, the Schmitt trigger generates a spike, and the neuron is reset using M_4 . The firing threshold of the neuron is determined by the Schmitt trigger circuit.

In one of the recent works, the integrate and fire effect was achieved by a neuron based on a single diffusive memristive device [65], illustrated in Fig. 5 (b). The diffusive memristor exhibits capacitive effect and a temporal behavior due to the doping of Ag nanoclusters between two electrodes of memristive material [64], [65]. In the application of such memristor as a neuron [65], it integrates the pre-synaptic signals, and when the memristor threshold is reached, the diffusive memristor changes its state and resistance of a memristor decreases causing a spike. The delay of a spike depends on the internal material properties and Ag doping in the diffusive memristor.

2) *Neuron model based on summing amplifiers and comparators*: Most of the ANN implementations use the neuron structures based on the summing amplifiers and comparators [67], [69], [70]. This model is usually used to represent threshold logic based linear neuron model. In most of the

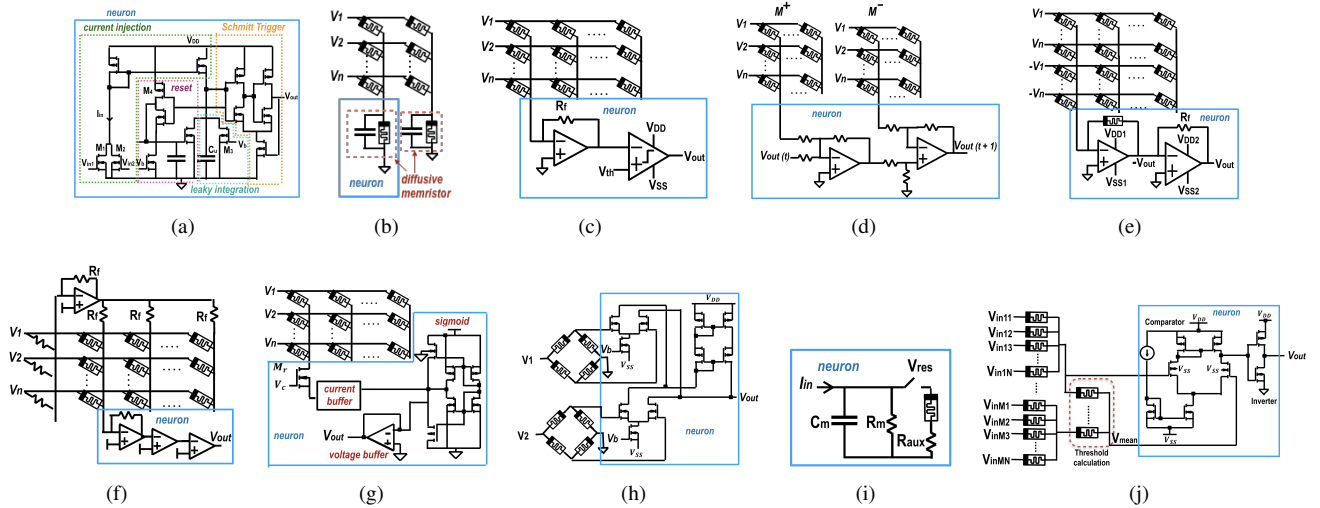


Fig. 5. Neuron cells: (a) Modified I&F neuron [68]; (b) memristor-based capacitive neuron [65]; variations of neuron models based on summing amplifier and comparator: (c) [69], [70], [32], (d) [71], (e) [33], and (f) [60]; (g) neuron models with sigmoid activation function [72], (h) neuron model for memristor-bridge architectures [56], [57]; (i) stochastic neuron [73]; and (j) HTM SP neuron [8].

cases, this structure is used for postsynaptic neurons, while presynaptic neurons have various configurations depending on the application of the architectures, or are not even shown in several research works. Different variations of such neurons are shown in Fig. 5 (c), Fig. 5 (d), Fig. 5 (e) and Fig. 5 (f).

Fig. 5 (c) represents the conventional summing and thresholding neuron configuration [69], [70]. The summing amplifier sums the input currents and outputs the equivalent voltage. The comparator output the spike or pulse (depending on the configuration of the circuit), when the amplifier output is above the threshold [69], [70]. Fig. 5 (d) shows a similar configuration of the output neuron with the summing amplifier combining the outputs from negative and positive memristive arrays and comparator circuit [71]. The other configuration is shown in Fig. 5 (e). The first amplifier is used to scale the output voltage and implement the sigmoid activation function, while the second unity gain amplifier inverts the output [33]. Fig. 5 (f) shows a neuron consisting of three amplifiers [60] used to sum the currents, invert the output and calculate the error, which allows updating the synapses.

3) Neuron models with different activation functions:

There are different ANN implementations which use various activation functions to implement the behavior of the neuron, such as sigmoid [72] and tangent [80]. One of such sigmoid-based neurons is shown in Fig. 5 (g) [72]. The neuron contains a sigmoid activation function with input current and output voltage and additional circuit to ensure the accurate performance and absence of loading effects. The currents from the memristive synapses are summed, and the current mirror is used to reduce the loading effect. The current is applied to the sigmoid activation function [81], and voltage buffer is used to normalize the sigmoid output. The voltage buffer is optional in this configuration.

4) *Neuron models for memristor bridge architecture:* The other possible implementation of the neuron is shown in Fig. 5 (h). These neurons correspond to the bridge synapse structure from [56], [57] and were proposed to be used only

with those synapses. In this neuron, the voltage weighted by the memristor bridge synapses is converted to the current using differential amplifiers [57]. Three transistors connected to the synapse represent voltage-to-current converter (VIC) acting as a current source. The neuron contains a self-biasing circuit to provide DC output current, an active load connected to all synaptic circuits which sum up the currents from all synaptic currents, and memristor load that converts output current into voltage. This circuit is used in various neural network architectures [57], [82]. Such configuration shows good performance for ideal simulations, however, if the circuit is constructed from the real memristors, the problems, such as switching response, switching time and connection issues of two memristors may occur. Also, if the number of connected synapses increases, the number of transistors in the neurons will increase significantly. Therefore, this is not the most efficient solution for very large architectures.

5) *Stochastic neurons:* In recent year, the exploration of the stochastic systems with added noise and memristor stochasticity gained the popularity. Such neuromorphic systems emulate the stochasticity in the cortex, where the biological noise helps the learning and information processing. In CMOS-memristive systems, stochasticity is introduced by ejecting the noise into the circuit. Either stochastic memristive synapses or stochastic neuron can be used for these purpose [73], [83]. One of the possible implementations of a stochastic neuron is shown in Fig. 5 (i). Memristor is arranged in parallel with original simple neuron circuit consisting of membrane resistor R_m and capacitor C_m [84]. The variable threshold of the memristor allows to randomize the firing threshold of the neuron and ensures random neuron spiking behavior. This stochastic memristor based neuron model tested for the architectures with 16 and 32 stochastic neurons is proposed in [73]. The stochastic neuron with memristor allows removing random number generator from the stochastic circuits. However, the application of such neurons for large-scale arrays is still questionable because of the size of the neuron due to the

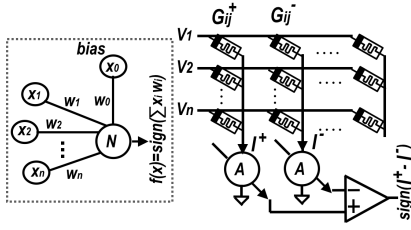


Fig. 6. One layer artificial neural network [61].

capacitor.

The application of the stochastic neurons for digits recognition problem is investigated in [73]. The accuracy that can be achieved is about 60 % for a system with stochastic neurons and 65 % for the stochastic synapses. The approach was tested for a small scale problem; however, it is mentioned that the 90 % of recognition accuracy can be achieved using 300 neurons or 235200 synapses. However, such architecture will have a large area and power consumption. The simulation of the system with stochastic memristive synapses in [85] allows achieving the recognition accuracy up to 82 % for MNIST database. The other stochastic spiking WTA network used for handwritten digits recognition with 78 % accuracy is shown in [83].

6) *HTM Spatial Pooler neuron*: The implementation of HTM neuron is not fully explored in terms of hardware realization. The implementation of inhibition phase of HTM Spatial Pooler (SP) that can be considered as a neuron cell is shown in Fig. 5 (j) [8]. The neuron consists of a comparator and inverter. This neuron is a part of modified HTM architecture, where the mean operation replaces the summation. The comparator performs the comparison of the mean voltage with the threshold, and the inverter normalizes the comparator output and produces the binary output. The variations of HTM neuron based systems are shown in [8], [86] and [87].

IV. NEUROMORPHIC ARCHITECTURES

A. Neural network architectures

There are different memristive neuromorphic architectures that can be used for edge computing applications. The summary of these architectures is shown in Table I. Also, there are several other memristive architectures proposed in the recent years, which are less common and not considered in this paper, such as Probabilistic Neural Networks [88], [89] and Binarized Neural Networks [90].

1) *One layer neural network with learning*: The structure of one-layer ANN with learning is similar to the feed-forward neural network but contains the learning phase. Learning can be performed using various learning rules, like Hebbian learning, backpropagation and different modifications of them. One of the implementations of one-layer ANN is shown in Fig. 6 [61]. The $2M Pt/TiO_{2-x}/Pt$ memristor synapses are used to ensure the negative sign of synaptic weights. The output is calculated as a binary activation function of a sum of all the synapses, which is equivalent to a perceptron [49]. The learning is performed using a perceptron learning rule, where the memristive synapses are strengthened or weakened

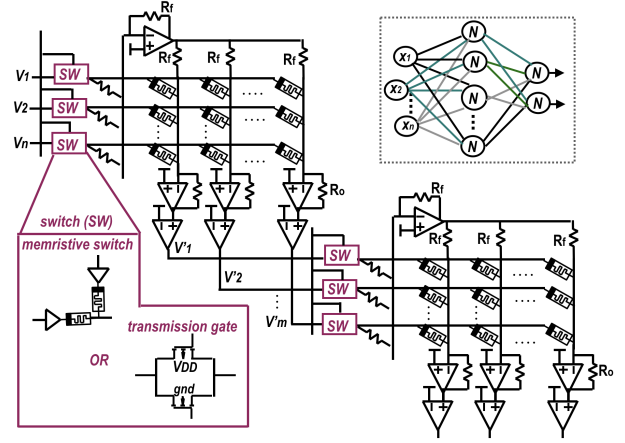


Fig. 7. Two layer neural network [59], [60].

depending on the desired output: $\Delta w_i = \pm \eta x_i (y_i - y_o)$, where y_i is an ideal output, y_o is a real output, η is a learning rate and x_i is an input. The architecture was tested for small-scale pattern classification problem.

The other implementation of one-layer ANN is proposed in [91]. The architecture is designed as an array of 2T1M synapses (Fig. 4 (e)). The performance was tested for handwritten digits recognition, and the obtained accuracy is approximately 83%. The implementation of one-layer ANN for face classification using single layer RRAM-based perceptron is shown in [54]. The architecture is constructed using $TiN/TaO_x/HfAl_yO_x/TiN$ 1T1M synapses (Fig. 4 (d)). The achieved average face recognition accuracy for Yale Face Database [92] is 88.08%.

2) *Two layer neural network*: The typical example of two layer neural network is a perceptron with a single hidden layer. Such architecture is shown in Fig. 7 [59], [60]. The architecture contains two crossbars with $Ag/AgInSbTe/Ta$ 1M synapses [60] and neuron cells shown in Fig. 5 (f). The control cell in the architecture contains either transmission gate [59] or memristive switch [60]. Both networks were tested for pattern recognition applications. The design is simulated for digits recognition problem with the accuracy up to 100% without noise [60].

Partially fabricated two-layer ANN with 64 input, 54 hidden and 10 output neurons shown in [93]. The 128×64 fabricated crossbar array was used in the network, while activation functions were implemented in software. The simulation was performed with rescaled images of size 8×8 pixels from MNIST database with the classification accuracy of 92%. The training was performed online, the update values for memristors have been calculated in software, according to backpropagation algorithm, and the corresponding update pulses were applied to the crossbar.

The other architecture for two-layer ANN proposed in [82] is based on 4M bridge synapses (Fig. 4 (f) and Fig. 5 (h)). The architecture is shown in Fig. 8. The architecture is similar to Radial Basis Network structure and consists of the artificial neurons with 7 CMOS transistors and memristor bridge synapses. The network with 432 inputs, 10 hidden neurons

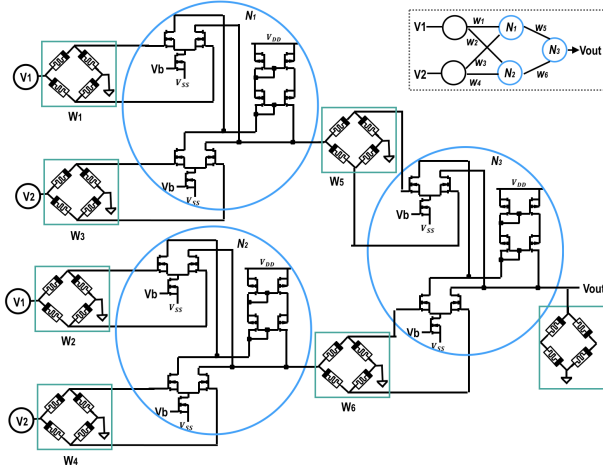


Fig. 8. Two layer neural network with memristor bridge synapses [82].

and 1 output neuron was tested for car detection problem using images of size 24×18 pixels. The results showed that the results obtained from circuit simulation are comparable with software simulation results. A similar approach is proposed in [57] with the implemented ANN is based on Random Weight Change (RWC) learning algorithm. The circuit implementation shows promising results in terms of processing time, which equals to $115ns$ in total for feedforward processing and the memristor programming.

3) *Deep Neural Networks*: Deep Neural Network (DNN) is a large class of the neural networks that consists of many cascaded layers and contains various activation functions between the layers. The number of layers in deep neural networks cause the scalability issues. Moreover, the application of memristive crossbars opens an opportunity to scale such networks staying at an acceptable level of power consumption. Therefore, memristor-based deep neural networks have been explored in the recent years. The architecture of memristive DNN is similar to two-layer neural networks but contains more crossbar arrays. The research work [33] explores the deep memristive convolutional neural network with 5 layers and reports the accuracy of 91.8% for MNIST handwritten digits classification. While [94] investigates the implementation of deep stochastic spiking convolutional 5 layer neural network with the MNIST classification accuracy of 97.84%, selecting the output class based on the largest number of output spikes produced by the output neurons. The energy consumption and on-chip area of this memristive network is 6.4 and 8 times smaller than in equivalent CMOS-based design, respectively.

4) *Cellular Neural Network*: The architecture of the cellular neural network (CeNN) is illustrated in Fig. 9. The architecture implies that the cells are connected only to the closest neighbor cells in the network. The first analog hardware implementation of cellular neural networks was proposed in the 1980s. The cells were designed with the capacitor, current source and resistive elements [96]. In contrast to this early design, the architecture of recently proposed CeNN is based on the memristive-CMOS circuits as shown in [56], [95]. The most commonly implemented memristive CeNN architecture

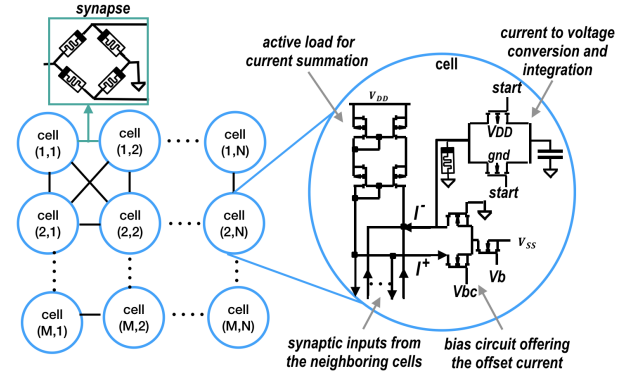


Fig. 9. Cellular neural network [95].

is based on 4M bridge synapses (Fig. 4 (f) and Fig. 5 (h)). The research work [97] illustrates the use of the memristor bridge circuit application with 5M synapses. This architecture is useful for the image processing tasks, such as edge detection [95], [98] and image filtering [56], [97]. The two dimensional CeNN architecture in the flux-charge domain is described in [99]. The CeNN can also be used for noise removal, extraction of horizontal lines and hole filling tasks.

5) *Convolutional Neural Network*: Convolutional Neural Network (CNN) is a machine learning algorithm based on a convolution operation that has been proven to be an efficient solution for various classification tasks, image recognition problems [100], [101] and video analysis [102]. Comparing to the software implementations of CNN, there are not many hardware implementations of CNN based on memristive circuits. Most of the hardware solutions for implementing CNN architecture are based on 1M memristive crossbar arrays or ReRAMs, while the processing units such as for implementing learning algorithm are digital [103], [104].

One of the hardware solutions for CNN is shown in Fig. 10 [33], [34]. The architecture is divided into feature extraction parts with convolution and sub-sampling (smoothing) layers and classification part. In CNN, the number of data features is reduced with the propagation through the network but the number of feature maps increases, which improves feature quality for inter-class discrimination. The convolution layer is followed with a fully connected multi-layered neural network that act as the classifier. The learning in such system is performed on software and the values of the memristors in each layer are programmed. The testing and classification are performed on hardware. In convolution layer, memristors represent the convolution filters and perform dot product calculation, similar to the fully connected layer. The output current from the crossbar is converted into voltage using the system with two amplifiers (as in Fig. 5 (d)). The number of memristors in each layer is determined by the initial size of images and number of required feature maps in this layer [34]. In [33] and [34], the size of the input images is 28×28 . In the first convolution layer, the image is filtered by 6 convolution filters producing 24×24 feature maps, while sub-sampling layer reduces the size of feature maps to 12×12 . In the second convolution layer the feature maps are filtered by 12

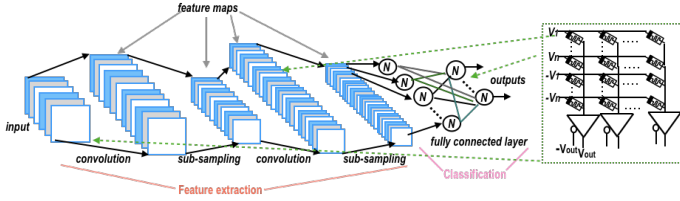


Fig. 10. Convolutional Neural Network [33], [34].

convolution filters producing feature maps of size 8×8 , which are reduced to the size of 4×4 in the second sub-sampling layer. The accuracy for handwritten digits recognition that can be achieved is 92% [33] and 94% [34], comparing to 98.92% of software simulation with MNIST database.

The research work [105] illustrates the memristive crossbar based accelerator for CNN implementation consisting of analog and digital components. In such systems, the analog components include only memristive crossbar; and most of the other components are digital. The power efficiency of such accelerator is 644.2 giga-operations per seconds (GOPS) per Watt (GOPS/W). The CNN accelerator based on the crossbar architecture with digital ReRAM is shown in [106] and [107]. The accuracy results are 98.3% and 91.4% for MNIST and CIFAR-10 databases, respectively. The area and power consumption of the system are $1.02mm^2$ and $6.3mW$. The system throughput is 792 (GOPS) and energy efficiency is 126 tera-operations per second (TOPS) per Watt (TOPS/W). The accuracy of CNN varies with the number of output feature maps from the convolution layer. The research work [108] illustrates that the implementation of CNN for MNIST character recognition using memristive crossbar has the on-chip area of $0.5033947mm^2$ and power consumption of $0.001785W$, which is more efficient in comparison to the implementation of CNN on the traditional RISC processor.

One of the most recent works in memristive convolutional filtering is illustrated in [109]. In this work, parallel vector matrix multiplication of array of size 128×64 is implemented. The current from all fabricated crossbar columns are read in parallel, which illustrates the speed of $1.64 TOPS$ for reading cycle. The power consumption of such crossbar is $13.7mW$, and the power efficiency is $119.7 TOPS/W$. Even though the image quality after convolution operation is worse comparing to software based convolution operation, memristive solution consumes 17 times less energy comparing to ASIC implementation. The recent work [110] illustrates the implementation of CNN in spike domain with digital memristor-based neuron using Time Division Multiplexing Access (TDMA) technique to reduce the number of required neurons. The classification accuracy of the network for handwritten digits recognition is 97%. However, the size and scalability of such network for the application on edge devices is an open problem.

6) *Spiking Neural Network*: In Spiking Neural Networks (SNN) the data signals are transmitted as spikes of a specific shape. This emulates the brain processing and is based on the particular spike events [42]. SNN focuses on the realization of plasticity rules and timing difference between pre- and postsynaptic spike. The spike based architectures are mostly

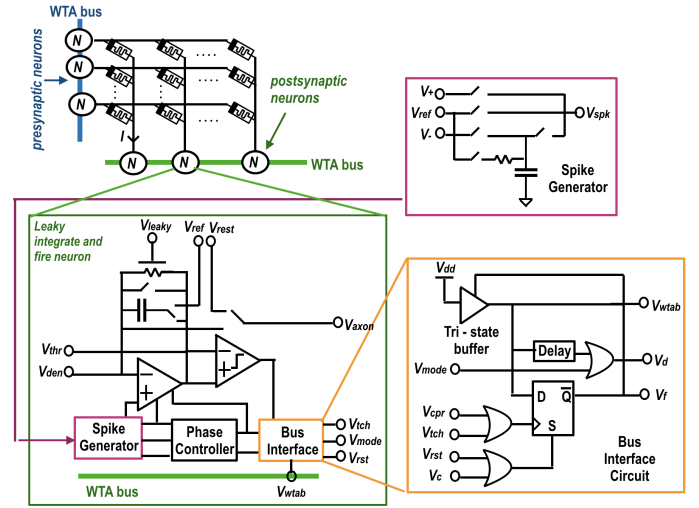


Fig. 11. Spiking Neural Network [74].

represented by Spike Timing Dependent Plasticity (STDP) implementation. STDP is based on biological concepts of presynaptic and postsynaptic impulses. The implementation of the neuromorphic architectures with STDP are based on the memristive crossbar arrays. The crossbar represents synapses and connected with the neuron models [111]. The possible implementation of such system is shown in [112]. Based on the correlation of the presynaptic and postsynaptic spikes, the synapse value between presynaptic and postsynaptic neurons represented by memristor is updated. Based on the postsynaptic neuron mode, the memristor is potentiated, depressed or stay unchanged. One of the advantages of SNN hardware implementation is that the power dissipation of such systems is smaller than in the pulse based systems. SNN can be used for handwritten digits recognition and letter recognition with the accuracy of up to 99% [113].

The basic SNN architecture is shown in Fig. 11; it consists of presynaptic neurons and postsynaptic neurons connected by 1M synapses [74], [114]. In most of the cases, the SNN is used with Winner-Takes-All (WTA) approach. One of such architectures for object position detection is introduced in [115]. Each input neuron corresponds to a particular position of the object, and the output neuron determines the exact position of the object based on the spiking frequencies. If the object is located between the input neurons, the spiking frequencies of the output neurons are proportional to the exact position of the object in the input neurons. Such position detector consisting of 5×5 neurons has maximum power consumption of $15.6\mu W$, which is about 70% less than equivalent CMOS design, and on-chip area of $6.1 \times 10^{-5}cm^2$.

The recent works introduce the stochasticity to SNN. The stochasticity implies the probabilistic behavior of neurons or synapses and represents the biological concept of the importance of neural noise during the information processing in the brain. In [84], the stochasticity is introduced to the simple Spiking WTA architecture shown in Fig. 13, where the output is determined by the first firing neuron from the output neurons. The simulation results from MNIST handwritten

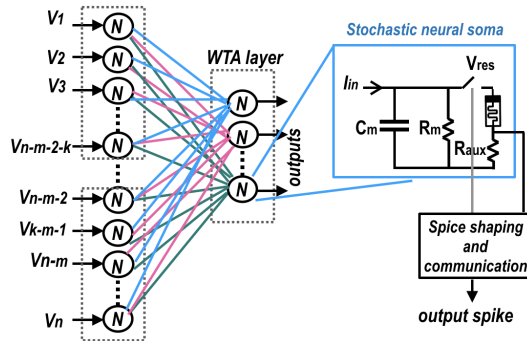


Fig. 12. Stochastic spiking neural network [84].

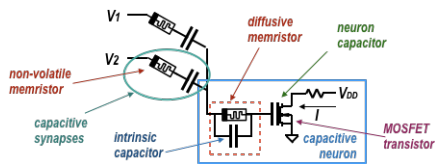


Fig. 13. Capacitive spiking neural network [116].

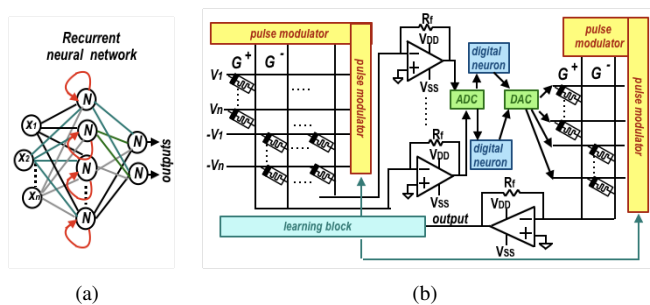


Fig. 14. (a) Recurrent neural network, and (b) Mixed Signal Implementation of one layer RNN [117].

digits recognition vary with the size of the layer of output neurons and reach 78.4% for 128 output neurons. The increase on the number of output neurons allows the network to capture more different patterns corresponding to the input data, which enhance the performance of majority voting procedure and allows to increase classification accuracy.

The alternative approach to implement SNN is a capacitive switching network presented in [116]. The resistive synapses are replaced with capacitive synapses concept. Capacitive synapses are based on non-volatile pseudo-memcapacitors formed by integrating non-volatile memristor in series with a capacitor. To form the capacitive neurons, the neuro-transistor is introduced, where dynamic pseudo-memcapacitors, formed by integrating recently proposed diffusive memristor with intrinsic capacitance [64], [65] in series with capacitor, are integrated onto the gate of a MOSFET. If the neuron is triggered by high capacitive state synapses, the post-synaptic neuron fires. The learning in such network is performed if presynaptic and postsynaptic neurons fire together causing the potentiation of low capacitance state in the synapses. Capacitive spiking network has an advantage of sneak-path free outputs.

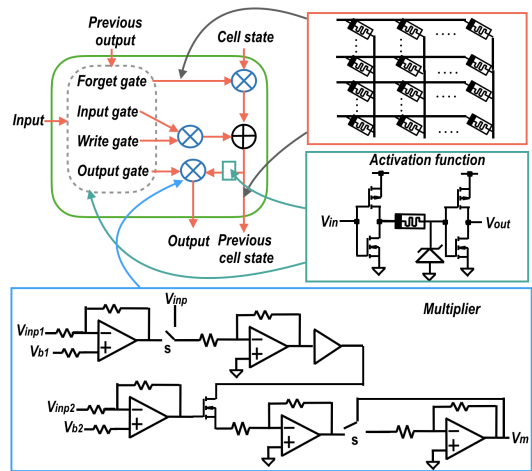


Fig. 15. Long Short Term memory [12].

7) *Recurrent Neural Network and Long Short Term Memory*: Recurrent Neural Network (RNN) is a neural network type, which involves the feedback calculation and the output of the layer effects the consequent outputs [118]. There are various architectures for RNN implemented in software; however, memristor-based hardware implementations of RNN is an open problem. There are several modifications of RNN, and simple RNN architecture is shown in Fig.14 (a). Fig. 14 (b) shows the implementation of one layer RNN with fabricated iron oxide memristive synapses [117], containing two parallel memristors representing positive and negative conductance, as in Fig. 4 (b). RNN design involves digital neuron, ADC and DAC, pulse modulator and learning block based on recursive least-squares algorithm.

The RNN architecture, especially in analog domain, has not been fully explored yet. Most of the works on memristive RNN focus on a mathematical analysis of system stability [119], [120]. The hardware implementation of memristive RNN is presented in [121]. The work illustrates analog implementations of the RNN using $0.5\mu\text{m}$ CMOS technology and applied for combinatorial optimization problems. Even though there are FPGA-based implementations of RNN [122], and some of the works show the possibility to integrate RNN with the memristive crossbar [121], the implementation of a full memristor based RNN architectures is an open problem. One of the main problems in analog implementations of RNN is the implementation of feedback and complexity of the architectures.

LSTM is a modification of RNN. One of the main features of LSTM is the feedback and selection of the information that effects future outputs. LSTM is based on modified dendritic threshold non-linear neuron model, where the output depends on the current input and previous outputs. The LSTM structure shown in Fig. 15 includes output gate, input gate, write gate and forget gate. These gates are responsible for how much the current output should be affected by the current inputs and previous outputs of LSTM. Memristor-based implementation of LSTM is proposed in [12]. The LSTM weights are presented as crossbars with 1M synapses, and the implemen-

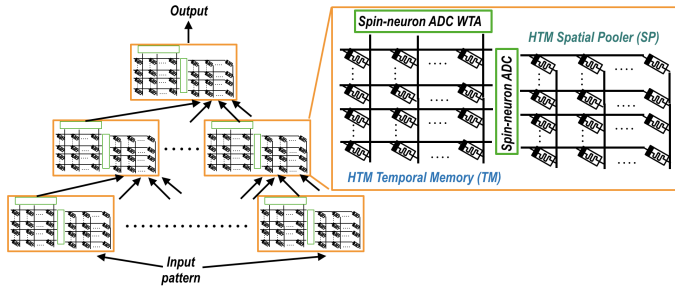


Fig. 16. Hierarchical Temporal memory [87].

tation of the activation function circuit that can be used as sigmoid or tangent is shown in Fig. 15. As LSTM algorithm requires multiplication, the analog multiplier is used. While the research work [12] shows the implementation of separate LSTM components, the full implementation of LSTM system is illustrated in [123] and [124]. Both systems are tested for the prediction of the number of airline passengers, and show the successful prediction of a trend, where LSTM system in [123] achieved the accuracy of 75%. The implementation RNN for edge inference with fabricated LSTM units based on memristive crossbar are shown in [13]. The implemented RNN consists of 15 LSTM units followed by fully connected layer. The LSTM is tested for prediction of the number of airline passengers and classification of an individual human by the persons gait, showing the precise prediction results and classification accuracy of 79%.

8) *Hierarchical Temporal Memory*: HTM is a machine learning algorithm and architecture mimicking the structure and functionality of human neocortex [10], [125]. HTM consists of HTM Spatial Pooler, which encodes the input patterns and produces sparse distributed representation of input data useful for visual data processing, and HTM Temporal Memory (TM), which can be used for prediction making [10]. Both HTM SP and HTM TM involve learning process. There are several CMOS-memristive hardware implementations of HTM proposed in recent years [8], [86], [87]. The mixed signal design of HTM is shown in [87], and the hierarchical structure of the proposed circuit is illustrated in Fig. 16. In this architecture, each level of HTM is presented by the memristive crossbar and spin-neuron devices are used as neurons for the processing. This architecture was used for MNIST handwritten digits recognition with the maximum accuracy of 95 % [87]. The architecture in [86] shows the alternative implementation of crossbar-based analog HTM SP circuit, which was tested for face recognition with AR [126] and speech recognition with TIMIT database [127] and achieved the accuracy of 86 % and 70%, respectively.

HTM architecture in [8] proposes the analog circuit level implementation of the modified HTM SP and HTM TM, inspired from HTM neuron shown in Fig. 3 (d). This architecture is shown in Fig. 17. In this system, HTM is used in combination with traditional supervised classification methods. Also, the implementation of HTM is modified to reduce the hardware level complexity. In this HTM system, HTM SP is used for encoding the input data patterns and

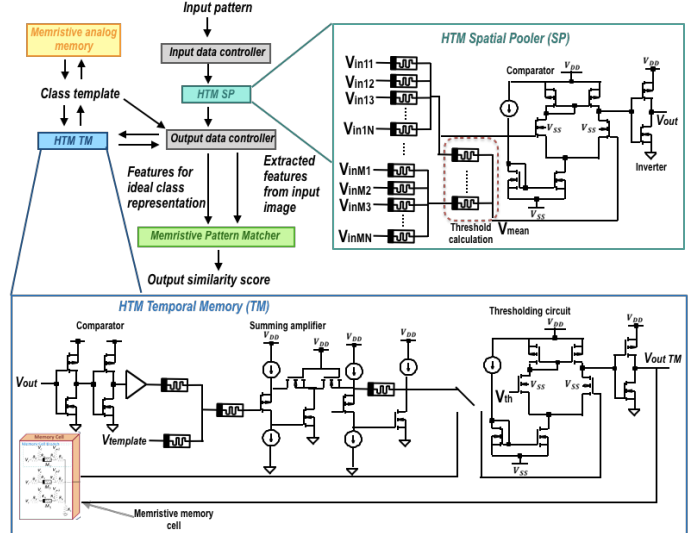


Fig. 17. Modified HTM Spatial Pooler and HTM Temporal Memory [8].

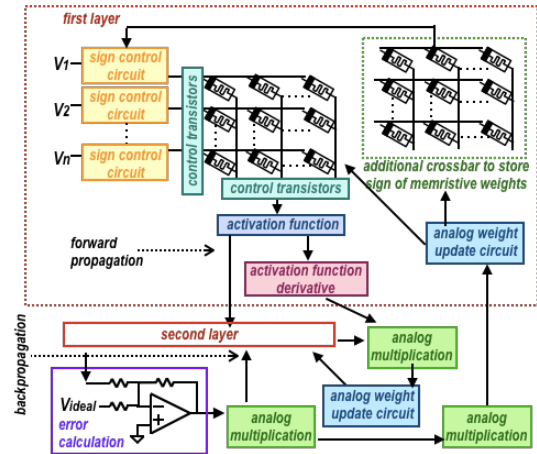


Fig. 18. Online backpropagation training architecture for memristive ANN [9].

presenting the inputs as sparse distributed binary patterns. In comparison with the traditional HTM algorithm, the HTM system uses HTM TM only for generation of the templates for all data classes stored in the memristive memory array. The classification is performed by the memristive pattern matcher, which compare the inputs processed by HTM SP with the ideal image templates. The synaptic weights are represented as separate memristors. The HTM SP part is based on the hardware implementation of HTM neuron shown in Fig. 5 (j). While the HTM TM part consists of the comparator and summing amplifier. The output of the HTM TM is used to update the training template stored in a memristive memory array [128]. The system is tested for face with AR database and speech recognition with TIMIT database, achieving the accuracy of 87% and 95%, respectively. There are several other HTM related works that propose different variations of hardware for memristive HTM [129], [130].

TABLE I
MEMRISTIVE NEUROMORPHIC ARCHITECTURES

| Architectures | Applications and simulation results | Scalability | Open problems | Drawbacks to improve for application in edge computing |
|----------------------|---|--|---|---|
| One layer ANN | handwritten digits recognition (83%) [91], face recognition (88.08%) [54] | Scalable with 1M devices | Investigation of the scalability of the system with 2T1M synapses | Investigation of the performance with real devices, processing speed, scalability, on-chip area and power dissipation for large scale systems, improvement of CMOS components |
| Two layer ANN | simple digits recognition (100%) [60] | Scalable with 1M devices, not scalable for bridge neuron | Investigation of the scalability of bridge neuron based systems and reduction of power dissipation of CMOS components | |
| Deep neural networks | various applications | Scalable with 1M devices | Investigation of the possibility of application for various problems, investigation of the effects of real memristors | Improvement of power dissipation and scalability issues |
| CeNN | image filtering | Not scalable | Investigation of the possibility to improve architecture for large scale simulations and to create the multilayer architectures | Investigation of the possibility to use with 1M devices to ensure the scalability of the system |
| CNN | handwritten digits recognition (94%) [34], | Partially scalable | Investigate the possibility of implementation of fully on-chip system without software part | As the number of layers is large, the scalability should be investigated |
| SNN | handwritten digits recognition(78.4%) [84], letter recognition [113] (99%) | Scalable | Investigation of the advantages over pulse-based systems and possibility to replace pulse based systems with spike based | Design of the scalable neurons producing spikes with small of chip area and power dissipation |
| RNN | pattern recognition | Scalable with 1M devices | Full circuit level design of the architecture, investigation of scalability and different applications | |
| LSTM | prediction making | Scalable | | |
| HTM | face recognition (98%) [86], [131], speech recognition (95%) [8], handwritten digits recognition (95%) [87] | Partially scalable | Implementation of full system performance, implementation of the exact algorithm for HTM SP and HTM TM, implementation of sequence learning in HTM TM | Improvement of CMOS components to ensure scalability |

B. Neural Network learning architectures

The learning process in the neural networks is important, especially for large-scale edge computing architectures. In memristive architectures for edge computing, the concept of online training is important [55], [132]. In most of the designs, the learning and online training of memristive architectures is performed on software. For example, partially fabricated neural network with online backpropagation training on software and online update of memristive weights of the crossbar is shown in [93]. However, it is important to ensure the scalability and low power dissipation in edge devices; therefore, separate software components and training units are not efficient for edge devices, and development of the architectures with online on-chip digital, mixed-signal and analog training architectures is important.

The online digital training and learning architectures based on the combination of memristive crossbars with digital training circuits for neural network implementation have been recently proposed in [66], [55], [31]. In [31], the digital training architecture for memristive DNN is proposed to accelerate the learning process and transfer it to hardware. The work [32] illustrates a mixed-signal design of neural network with analog neurons and digital error calculation and on-chip training.

For near sensor processing, it is essential to use analog systems that can be easily integrated with analog sensors

without additional stages of analog to digital and digital to analog conversion. Several works investigate the analog learning circuits for neural networks [72], [9], [60] and HTM [8]. In the implementation of backpropagation shown in Fig. 7 [60], the errors from the output neurons in the second layer are propagated back, and the memristors of the second and first neural network layer are updated sequentially. The memristors of the layer, which is not currently updated, are isolated by the memristive switch. The amount of the update value is proposed to be calculated on FPGA or using Look-Up Table (LUT).

The other recently proposed training architecture is illustrated in Fig. 18 [9], showing the complete hardware implementation of the calculation of backpropagation of error with the derivative of the activation function, relevant multiplication circuits, control transistors and analog weight update circuit. Also, the research work [9] illustrates the application of analog backpropagation circuit for different analog learning architecture, such as Multiple Neural Network (MNN) containing several neural networks processing different types of data and ANN decision layer, Binary Neural Network (BNN) based only on two state memristors, DNN, LSTM and HTM. Even though, several memristive analog implementations of neural networks has been proposed recently, the optimization and testing of fully analog learning systems with control circuitry without digital processing is still an open problem.

In the online training, one of the main issues of the learning process in memristor-based architectures is the update speed of the memristive weights. To update weights in a memristive crossbar, different update techniques can be used. The memristive synapses containing only memristors (1M,2M) and memristive synapses with transistors (1T1M and 2T1M) can be updated one at a time, which is a slow process. In this scheme, 1T1M and 2T1M allow to disconnect the memristors which are not involved in the update process completely and eliminate the leakage currents and effect on those memristors. To speed up the learning process, memristors in a crossbar can be updated in 2 steps: 1) update all memristive weights requiring the change from R_{ON} to R_{OFF} , and 2) update the others requiring the change from R_{OFF} to R_{ON} [133]. This method can be more efficient for the small crossbars with negligible leakage current and for modular crossbar approach, which is proposed to reduce the leakage currents in the memristive crossbar by dividing a large crossbar into smaller sub-crossbars [134], where all sub-crossbars can be updated in parallel reducing the training time.

V. DISCUSSION

This section includes the discussion of the advantages of memristive neuromorphic architectures, challenges that may occur during the simulation and implementation of the real system and open problems that should be addressed for efficient implementation and integration of neuromorphic architectures into the edge devices. In the simulation of such large neuro-memristive networks, the selection of memristor model is one of the challenging tasks, which is discussed in Appendix A.

A. Advantages of memristive architectures

The main advantages of the memristor-based systems for edge computing applications are the small on-chip area, low power dissipations, and scalability of the memristor-based systems. Therefore, the memristor circuits are a promising solution for edge-computing devices, where the computation is performed on the device without sending information into the cloud.

1) *Push from market and users*: The increased number of edge devices in Internet of things and Cyber Physical System frameworks is driven by the needs from the users for applications such as for gaming, object detection, augmented reality, artificial intelligence, video analytic, and mobile computing [135], [136]. This demands devices and chips that consume low energy, smaller area, and can provide higher computational capacity. The memristive architectures is envisaged to have this potential to achieve these objectives promoting more than Moore's law integration, and emerging intelligent applications [20].

2) *On-chip area and power dissipation*: The advantages of the implementation of memristive circuits include the significant reduction of on-chip area and power dissipation. In several systems, memristor is proposed to be used instead of resistors due to the small on-chip area and low power dissipation. For example, in comparison to CMOS-based design, for the

memristive CAM array design, on-chip area and average power consumption are reduced by 45% and 96%, respectively [137]. The area of memristive devices varies based on the used materials and the required resistive levels. The area of memristive devices of various materials can vary from micron to sub-10 nm depending of the required device properties [47], [52], [113], [138].

3) *Scalability*: The application of memristive devices allows scaling the systems because memristor does not exhibit leakage current problems, comparing to transistors and resistors. One of the most efficient solutions is scalable memristive crossbar structures. However, large crossbars can exhibit sneak path problems and the small variability of crossbar outputs. As a solution to this problem, the scalability of the memristive circuits and arrays can also be achieved by dividing the large memristive arrays into smaller sub-arrays [139]. Other well known solutions are to use selector devices along with memristors as outlined in previous sections, which however increases the cell area.

B. Major issues, open problems, and future work perspective

Even though there are a lot of benefits of memristor-based systems for edge computing applications, the research field of memristive circuits is not mature enough for commercial chip design solutions. Therefore, there is many drawbacks and open problems that can be investigated in future, such as compatibility issues, unstable switching behavior, limitations in the range of resistance and number of resistive levels, the complexity of fabrication of memristive systems and various issues of implementation of large-scale complex systems.

1) *Memristor materials and compatibility issues*: One of the major issues of the memristive circuits based design is the compatibility of memristive elements with the CMOS technology and fabrication issues. Several memristive devices are proven to be compatible with the CMOS fabrication process [137], [140]. While TiO_{2-x} memristors were quite popular, there are other growing list of memristors based on materials such as HfO_x , TaO_x , MoO_x , $La_{1-x}Sr_xMnO_3$, $InGaZnO$ [141], organic memristors with electrografted redox thin film [142], ferroelectric tunnel memristors (FTM), $Ge_2Sb_2Te_5$ (GST) memristors [132], SiO_x [143], SiN_x [141] and $Pr_{0.7}Ca_{0.3}MnO_3$ (PCMO) [63]. As the memristor technology is only at early stages of development, the properties, stability issues, switching behavior and compatibility with CMOS devices of various memristive elements and selection of most stable material stack is an open problem.

2) *Variability in switching behavior*: The variability issues are common in the memristive devices due to the immaturity of the memristive technology. The switching behavior of the memristive devices may vary, which affects the performance accuracy of many architectures [144]. Even though most of the memristor models used for simulations illustrate the ideal switching behavior, the real devices show the variability in switching behavior. Several works investigate the probability of switching of the memristive devices and apply this property in the stochastic systems [145]. While the stochasticity in switching may be useful for some systems, the effects of this

behavior on various neuromorphic architectures and learning systems have not been investigated yet. While, there have been works that have shown that the use of learning can compensate for variability at system level in digital neural architectures, implementation of learning algorithms with memristors for analog neural network remains a challenging problem. In addition, the effect on the learning process and training speed should also be explored. It should be also noted that there are several memristor devices proposed in the last decade, device to device variability is high and large majority of them are still in its infancy for industrial use.

3) *Range of resistance and number of stable states in memristive devices:* According to the material and physical properties, different memristive devices can be programmed into different ranges of resistance different and the number of stable resistive states. In most of the cases, the neuromorphic architectures are designed for a specific range of resistance and do not take into account the restricted number of resistive levels when simulating the overall system. In the real devices, depending on the material and fabrication process by changing the width of active layer, these parameters can vary, and the number of resistive states is finite. The recent research works show the memristive devices can achieve up to 64 stable resistive states [109]. One of the solution to increase the number of resistive level is to use parallel and series combination of memristive devices, which also implies that the issue of memristor interconnection should be considered. The issue of limited number of the resistive levels can be mitigated adding the additional circuits and components, however this increases complexity, number of components and power dissipation.

From the device perspective, the open problems include the investigation of the possibility to improve the number of resistive states and the investigation of possible materials that can be used for such purposes. From a mathematical modeling perspective, the model of the memristor incorporating the limited number of stable resistive states and non-linearity of the switching between different states that reflect a realistic memristor is still an open problem. From circuit and system design perspective, it is essential to consider the limited number of stable resistive states in the design and investigate the effect of this issue on the overall system performance.

4) *Endurance of the memristor:* Lifetime and reliability of memristive devices is a subject for the investigation, as there is a large number of memristive materials, in which endurance properties may vary. For example, [146] reports that TiO_x and TaO_x devices have an endurance of 10^5 and 10^9 cycles for $1\mu s$ applied voltage pulses, respectively. The endurance and reliability of memristive devices depend on process variability, including device-to-device and cycle-to-cycle variations, and endurance degradation referring to limited number of update cycles [147]. Cycle-to-cycle variability depends on the material of a memristor, while device-to-device variability refers to time-varying device stability depending on the manufacturing process and operation parameters, such as voltage, temperature and duration of applied voltage pulses [147]–[149].

In the edge computing architectures, especially involving the learning and training process, the lifetime of the memristor and number of possible update cycles is critical. For example,

the online learning process to train simple two layer ANN for simple XOR problem requires 5000 training iterations [9], and involves the continuous update of the memristive synapses. Moreover, to achieve a high-performance accuracy of the neural network, usually the learning rate is decreased, which leads to the requirement to increase the number of update cycles [9]. Therefore, it is important to investigate the endurance, reliability and lifetime limits of various memristive devices.

5) *Integration with CMOS devices and CMOS issues:* Considering the current trends in the technology market, it will be impossible to avoid the integration of the memristive devices into the CMOS architectures. Considering the importance of the implementation of the read and write circuits for the memristive devices, which are mostly based on the CMOS transistors, the number of CMOS devices per chip will be increased with the increase of the size of memristive architectures, primarily when the synapses or neurons in the neuromorphic architectures are based on hybrid CMOS-memristive designs.

Even though the memristor is a two terminal vertical element [150], which ensures the reduction of the on-chip area of the CMOS-memristive circuits, the fabrication process of the complex neuromorphic architectures may still be difficult. In a complex multilayer structure, where the memristive arrays are combined with CMOS circuits, the fabrication temperature is a critical issue. In combination with CMOS devices, high deposition temperature can damage the devices, while low temperature cannot guarantee the reliable connection between the elements. This also may increase the cost of such memristor-based architectures and systems. There is a variety of materials that exhibit memristive behavior; however, not all of them can be used for the fabrication of the complex architectures. The fabrication issues should be considered during the design stage and selection of the memristive elements. In the recent years, the successful integration of memristive devices into CMOS architectures is performed using Back End Of Line (BEOL) process and building a layer of memristors on top of the existing chip [151].

In addition, the increase in the number of CMOS devices on a chip, especially for such complex architectures as neural networks, leads to high power consumption. To avoid this issue, the size of CMOS devices should be decreased leading to lower supply voltages. As it is impossible to decrease the size of the CMOS devices further and maintain an accurate and precise performance of the device at the same time, the replacements of the CMOS devices, such as FinFET devices, should be further investigated and used in the memristive circuits.

6) *Implementation of large scale systems:* The investigation of complex multilayer architectures and systems is essential to ensure the scalability and accuracy of edge computing devices. Most of the recent works representing the complex multilayer systems are digital and based on Field Programmable Gate Arrays (FPGA). However, for edge devices that are restricted in terms of area and power consumption, FPGA is not an efficient solution. The number of complex mixed-signal and analog implementations of the neuromorphic systems and

architectures is limited. There are plenty of implementations of simple neural networks, such as perceptron and feedforward neural network, which proves the concepts and illustrates a solution of a particular problem using a particular database. However, more complex and generalized systems have not been investigated yet, and it is important to consider scalability and performance issues of multilayer systems. Also, in a full chip design of a complex system, it is important to consider the interconnection of memristive circuits with the other elements and complexity of datapath. These are design specific and application dependant issues. For signal processing in analog domain, the interconnection of elements can introduce parasitics having a significant impact on a system performance, comparing to digital signal processing, where the effect of signal integrity issues can be easily mitigated. The interconnect networks for memristive crossbars are studies in [152], [153]. As a memristor is a vertical device and the number of layers in the deep learning systems can be substantial, the possibility of implementation of vertical on-chip systems can also be investigated.

VI. CONCLUSION

In this paper, we presented an overview of a range of neuro-memristive circuits and architectures that is suitable to be developed as integrated circuit chips in edge computing devices. The pressing hardware issues and challenges involving emerging memristive circuits are presented. The growth of Internet of things and its growing impact on applications for drives the need to have smarter and faster computing in edge devices. Neuro-memristive architectures aims to emulate algorithms such as that based on neural networks and information processing mechanisms in human brain. The ability to (1) have lower on-chip area and power requirements, and (2) incorporate analog dot-product computing with memristive arrays, enables a highly efficient and scalable implementation possibility for on-chip neural networks. While these architectures can be a promising solution for efficiency and energy issues of edge devices, various challenges and drawbacks should be considered during the design to make their architectures applicable for edge devices. The open problems include various memristive device issues, the ability of integration and implementation of complex systems.

APPENDIX A SELECTION OF MEMRISTOR MODEL

To move from theoretical designs and simulations of the neuromorphic architectures to the implementation of the real chips and integration on the neuromorphic designs into the existing sensors, it is important to consider side effects, nonlinearities and drawbacks of the memristive circuits during the simulation process [154]. The selection of the memristor model can affect the simulation results significantly. The ideal memristor models will not consider non-linear effects of the real implementation, and the simulation results will not be reliable. While the memristor models that are not designed for large-scale simulations may cause the simulations errors and non-convergence issues in the SPICE simulation of large

architectures. Table II illustrates the most commonly used memristor models and their characteristics. More comprehensive review and consideration of the other memristor models is provided in [155]–[157].

1) *Early linear approximations and equivalent circuits:*

One of the earliest works on linear approximations and equivalent circuits for current and voltage-based memristor models is proposed in [166]. The circuits introduce the basic memristor concepts and are not used in the recent memristive architectures and systems. The equivalent circuit based memristor macro models are shown in [167]–[169]. These models are rarely used in the large-scale system simulation due to the complexity and lack of consideration of non-linearity and physical parameters of real devices.

2) *Linear memristor models:* The other major class of the memristor models is linear ideal models. Linear memristor model emulates the switching behavior of the devices and does not consider the effects of electric field on the device performance. The simplest linear memristor model is shown by Pickett at al. in [158]. This model is based on drift mechanism of ionized dopants and emulates TiO_2 memristor [84]. The linear relationship between the voltage and current in the memristor can be described as $v(t) = (R_{ON} \times x(t) + R_{OFF}(1-x(t))) \times i(t)$, where $x(t) = w(t)/D$ and D is a width of the device and $w(t)$ is a width of a doped region at a particular time [170]. The linear window function can be shown as: $f(w) = w(Dw)/D$ [159]. Even though this memristor model is frequently used in the simulations of neuromorphic circuits [171], it does not show various effects of non ideal behavior of real memristive devices.

3) *Nonlinear memristor models:* In the real device, the drift, diffusion, and thermophoresis due to ionic motion cause the nonlinear relationship between memristor current and voltage as well as nonlinear dynamical switching behavior [172]. In comparison to the ideal linear memristor models, like as Pickett model [158], that was used earlier to simulated the memristive architectures and prove the concept of the design of various neuromorphic architectures, the recent research works focus on memristor model containing nonlinearity effects. It is vital to consider non-idealities of the memristive devices because the memristive technology is not mature yet. The lack of stability makes the nonlinearity factor to be relevant to investigate, primarily when the large-scale simulations are performed.

One of the known nonlinear memristor models is Joglekar's model that allows controlling a non-linearity windowing function [161]. The main parameters causing non-linearities are W/D ratio and p , where W are the actual width, D is a width of the thin film, and p is a parameter of the window-function for modeling of nonlinear boundary conditions. The Joglekar's window function is represented as $f(x, p) = 1 - (2x - 1)^{2 \times p}$. According to this equation, the parameter p is responsible for the linearity of the memristor model that increases with the increase of p . The example of the application of this model is the CeNN architecture shown in [99].

The memristor shown in research work [162] is used in several neuromorphic architectures. The macro model is based on the study of the behavior of the TiO_2 memristor illustrated

TABLE II
COMPARISON OF THE MEMRISTOR MODELS

| Memristor model | Description | Linearity | Consideration of physical parameters of the memristor | Application for large scale simulations |
|---|--|-----------------------|---|--|
| Linear dopant drift models [158], [159] | Emulate the switching behavior of the devices and do not consider the effects of electric field and nonlinearities | Linear | partially considered | less computationally complex than non-linear models; however can only be used for a proof of concept [156] |
| Nonlinear dopant drift models [160]–[162] | Models with different window functions and consider the non-linear switching behavior | Non-linear | not considered | reduced simulation speed due to the complexity of window function |
| TEAM model [163] | Generalized model containing various window functions, nonlinear switching and effect of physical parameters | Non-linear | considered | difficult to use in extremely large arrays due to the complexity |
| Modified Biolek's models [156], [164] | Modification of the existing models designed for simulation improvement | Linear and non-linear | partially considered | can be used for large scale simulations without numerical problems and convergence issues |
| Data driven simplified model [165] | Model contains a window function allowing the derivation of a resistive state time-response expression for constant bias voltage | Non-linear | considered | includes data driven parameters and can be used for large scale simulations without convergence issues |

in [159]. The model allows modifying the nonlinear boundary conditions that are not considered by the simplified linear memristor models. The model is based on the modification of non-linear window function from [161] demonstrating nonlinearities caused by non-linear dopant drift. The main parameters causing non-linearities are i , W/D ratio and p , where i is a current flowing through the memristor. The window function is represented as: $f(x, i, p) = 1 - (x - \text{step}(-i))^{2 \times p}$, where $\text{step}(-i) = 0$ for $i < 0$ and $\text{step}(-i) = 1$ for $i \geq 0$. The window function is involved into the calculation of the resistance value of the memristor and the speed of the movement of the boundary between the doped and undoped regions of the memristor, which determines how fast the resistive state of a memristor changes. The main difference between the Joglekar's and Biolek's memristor models [162] is the ability of Biolek's model to reversely change the memristance after a reaching one of the resistance boundary [63].

The non-linear model that can be adjusted and scaled is proposed in [160]. This window function for this model is the following: $f(w) = j(1 - [(w - 0.5)^2 + 0.75]^p)$, where j represents a control parameter to specify highest value of window function [160], [170].

Even the memristor models proposed in [161], [160] and [162] contains the nonlinear switching behavior of the memristor, the nonlinearities of the device, parasitic effects, leakages and other physical imperfections are not considered. The physical imperfections of the device are considered in [173] and [163].

The memristor model that is used in many neuromorphic architectures is ThrEshold Adaptive Memristor (TEAM) model [95]. The model is proposed in [163]. The model represents a generalized solution for Simmons Tunnel Barrier model [173], [174] that is complicated and designed for a particular memristor type [170]. This model illustrates a generalized approach and can be adjusted for the behavior of different memristive devices and different window functions considering various physical effects [173]. The model is used in different system level simulations and is proven to be useful for fast digital systems [121].

4) *Memristor models for large-scale simulations*: Most of the architectures are realized on memristive crossbars and the amount of processed data is significant, especially in edge computing when the data is not sent to the cloud. While the proof of the concept and overall ability of the system to perform a certain task can be tested using ideal memristors, the real performance of the memristive systems require more accurate non-linear memristor models. Therefore, one of the most important aspect in memristor modeling is to take into account the non-linearity problem and physical effects. However, the large-scale simulations of memristive systems cause various numerical problems and make it impossible to check the performance of the large-scale system. The large number of internal equations in the models and the mathematical form of those equations, especially the ones incorporating non-linear dynamics of memristive devices, can cause the problems of data overflow and convergence issues [156]. Therefore, it is important to find a trade-off between the accuracy of memristor models and computational complexity.

The modification of memristor model that can be used for large-scale simulations is shown in [164] and [156]. The model does not contain the window functions and allows to avoid different numerical problems and non-convergence issues [156]. The research work [156] proposes the modification of complex physicalphenomenological nonlinear models appropriate for large-scale simulations of multilayer architectures for edge computing. The other recent model suitable for large scale simulations is shown in [165]. This model simplifies the data fitting process, introduces a window function allowing the derivation of a resistive state time-response expression for constant bias voltage, and provides the possibility to perform computationally efficient simulations of the designed architectures for more realistic conditions. The model is data driven and provides the example of fitting parameters for TiO_x and TaO_x devices.

REFERENCES

- [1] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.

- [2] T. R. Sheltami, E. Q. Shahra, and E. M. Shakshuki, "Fog computing: Data streaming services for mobile end-users," *Procedia computer science*, vol. 134, pp. 289–296, 2018.
- [3] C. Panetta, "Top trends in the gartner hype cycle for emerging technologies. 2017," *Enterprises should explain the business potential of blockchain, artificial intelligence and augmented reality*, 2017.
- [4] M. Gusev and S. Dustdar, "Going back to the roots x2014: the evolution of edge computing, an iot perspective," *IEEE Internet Computing*, vol. 22, no. 2, pp. 5–15, Mar 2018.
- [5] G. Premsankar, M. D. Francesco, and T. Taleb, "Edge computing for the internet of things: A case study," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1275–1284, April 2018.
- [6] G. Chakma, M. M. Adnan, A. R. Wyer, R. Weiss, C. D. Schuman, and G. S. Rose, "Memristive mixed-signal neuromorphic systems: Energy-efficient learning at the circuit-level," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 1, pp. 125–136, March 2018.
- [7] L. Chua, V. Sbitnev, and H. Kim, "Hodgkin–huxley axon is made of memristors," *International Journal of Bifurcation and Chaos*, vol. 22, no. 03, p. 1230011, 2012.
- [8] O. Krestinskaya, T. Ibrayev, and A. P. James, "Hierarchical temporal memory features with memristor logic circuits for pattern recognition," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. PP, no. 99, pp. 1–1, 2017.
- [9] O. Krestinskaya, K. N. Salama, and A. P. James, "Learning in memristive neural network architectures using analog backpropagation circuits," *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp. 1–14, 2018.
- [10] O. Krestinskaya, I. Dolzhikova, and A. P. James, "Hierarchical temporal memory using memristor networks: A survey," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 5, pp. 380–395, Oct 2018.
- [11] D. George and J. Hawkins, "Hierarchical temporal memory: Concepts, theory and terminology," Tech. Rep., 2006. [Online]. Available: <http://www-edlab.cs.umass.edu/cs691jj/hawkins-and-george-2006.pdf>
- [12] K. Smagulova, O. Krestinskaya, and A. P. James, "A memristor-based long short term memory circuit," *Analog Integrated Circuits and Signal Processing*, pp. 1–6, 2018.
- [13] C. Li, Z. Wang, M. Rao, D. Belkin, W. Song, H. Jiang, P. Yan, Y. Li, P. Lin, M. Hu *et al.*, "Long short-term memory networks in memristor crossbars," *arXiv preprint arXiv:1805.11801*, 2018.
- [14] AppleInsider, "iphone xs a12 bionic chip features 7nm design, next-gen neural engine," Sep 2018. [Online]. Available: <https://appleinsider.com/articles/18/09/12/iphone-xs-a12-bionic-chip-features-7nm-design-next-gen-neural-engine>
- [15] "The world's first 7nm process mobile ai chipset," Nov 2018. [Online]. Available: <https://consumer.huawei.com/en/campaign/kirin980/>
- [16] C. D. Schuman, T. E. Potok, R. M. Patton, J. D. Birdwell, M. E. Dean, G. S. Rose, and J. S. Plank, "A survey of neuromorphic computing and neural networks in hardware," *arXiv preprint arXiv:1705.06963*, 2017.
- [17] A. B. Kahng, "Scaling: More than moore's law," *IEEE Design & Test of Computers*, vol. 27, no. 3, pp. 86–87, 2010.
- [18] M. T. Bohr and I. A. Young, "Cmos scaling trends and beyond," *IEEE Micro*, vol. 37, no. 6, pp. 20–29, 2017.
- [19] X. S. Hu, "A cross-layer perspective for energy efficient processing: from beyond-cmos devices to deep learning," in *Proceedings of the 2018 on Great Lakes Symposium on VLSI*. ACM, 2018, pp. 7–7.
- [20] F. Balestra, M. Graef, B. Huizing, Y. Hayashi, H. Ishiuchi, T. Conte, and P. Gargini, "Executive summary," *International Roadmap for Devices and Systems*, 2017.
- [21] X. Wang, Y. Chen, H. Xi, H. Li, and D. Dimitrov, "Spintronic memristor through spin-torque-induced magnetization motion," *IEEE electron device letters*, vol. 30, no. 3, pp. 294–297, 2009.
- [22] D. Kuzum, R. G. Jeyasingh, B. Lee, and H.-S. P. Wong, "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing," *Nano letters*, vol. 12, no. 5, pp. 2179–2186, 2011.
- [23] C. Ho, E. K. Lai, and K. Y. Hsieh, "Programmable resistive ram and manufacturing method," Sep. 29 2009, uS Patent 7,595,218.
- [24] K.-H. Kim, S. Gaba, D. Wheeler, J. M. Cruz-Albrecht, T. Hussain, N. Srinivasa, and W. Lu, "A functional hybrid memristor crossbar-array/cmos system for data storage and neuromorphic applications," *Nano letters*, vol. 12, no. 1, pp. 389–395, 2011.
- [25] F. T. Hady, A. Foong, B. Veal, and D. Williams, "Platform storage performance with 3d xpont technology," *Proceedings of the IEEE*, vol. 105, no. 9, pp. 1822–1833, 2017.
- [26] H. Jiang, C. Li, R. Zhang, P. Yan, P. Lin, Y. Li, J. J. Yang, D. Holcomb, and Q. Xia, "A provable key destruction scheme based on memristive crossbar arrays," *Nature Electronics*, vol. 1, no. 10, p. 548, 2018.
- [27] A. K. Maan, D. A. Jayadevi, and A. P. James, "A survey of memristive threshold logic circuits," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 8, pp. 1734–1746, 2017.
- [28] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.
- [29] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3590–3605, 2016.
- [30] G. Chakma, N. D. Skuda, C. D. Schuman, J. S. Plank, M. E. Dean, and G. S. Rose, "Energy and area efficiency in neuromorphic computing for resource constrained devices," in *Proceedings of the 2018 on Great Lakes Symposium on VLSI*. ACM, 2018, pp. 379–383.
- [31] M. Cheng, L. Xia, Z. Zhu, Y. Cai, Y. Xie, Y. Wang, and H. Yang, "Time: A training-in-memory architecture for memristor-based deep neural networks," in *2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2017, pp. 1–6.
- [32] R. Hasan, T. M. Taha, and C. Yakopcic, "On-chip training of memristor crossbar based multi-layer neural networks," *Microelectronics Journal*, vol. 66, pp. 31–40, 2017.
- [33] C. Yakopcic, M. Z. Alom, and T. M. Taha, "Memristor crossbar deep network implementation based on a convolutional neural network," in *2016 International Joint Conference on Neural Networks (IJCNN)*, July 2016, pp. 963–970.
- [34] —, "Extremely parallel memristor crossbar architecture for convolutional neural network implementation," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 1696–1703.
- [35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [36] H. Abunahla and B. Mohammad, *Memristor Technology: Synthesis and Modeling for Sensing and Security Applications*. Springer, 2018.
- [37] J. Rajendran, O. Sinanoglu, and R. Karri, "Regaining trust in vlsi design: Design-for-trust techniques," *Proceedings of the IEEE*, vol. 102, no. 8, pp. 1266–1282, Aug 2014.
- [38] J. Rajendran, G. S. Rose, R. Karri, and M. Potkonjak, "Nano-ppuf: A memristor-based security primitive," in *VLSI (ISVLSI), 2012 IEEE Computer Society Annual Symposium on*. IEEE, 2012, pp. 84–87.
- [39] A. Mazady, M. T. Rahman, D. Forte, and M. Anwar, "Memristor pufa security primitive: Theory and experiment," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 5, no. 2, pp. 222–229, 2015.
- [40] G. S. Rose, J. Rajendran, N. McDonald, R. Karri, M. Potkonjak, and B. Wysocki, "Hardware security strategies exploiting nanoelectronic circuits," in *Design Automation Conference (ASP-DAC), 2013 18th Asia and South Pacific*. IEEE, 2013, pp. 368–372.
- [41] P. Tino, L. Benuskova, and A. Sperduti, "Artificial neural network models," in *Springer Handbook of Computational Intelligence*. Springer, 2015, pp. 455–471.
- [42] S. Carrillo, J. Harkin, L. J. McDaid, F. Morgan, S. Pande, S. Cawley, and B. McGinley, "Scalable hierarchical network-on-chip architecture for spiking neural network hardware implementations," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 12, pp. 2451–2461, Dec 2013.
- [43] R. D. Cază, M. D. Humphries, and B. S. Gutkin, "Dendrites enhance both single neuron and network computation," in *The Computing Dendrite*. Springer, 2014, pp. 365–380.
- [44] J. Hawkins and S. Blakeslee, "On intelligence. 2004," *New York St. Martins Griffin*, pp. 156–8.
- [45] J. Hawkins and S. Ahmad, "Why neurons have thousands of synapses, a theory of sequence memory in neocortex," *Frontiers in neural circuits*, vol. 10, p. 23, 2016.
- [46] R. Rojas, *Neural networks: a systematic introduction*. Springer Science & Business Media, 2013.
- [47] Z. Wang, W. Zhao, W. Kang, Y. Zhang, J. O. Klein, and C. Chappert, "Ferroelectric tunnel memristor-based neuromorphic network with 1t1r crossbar architecture," in *2014 International Joint Conference on Neural Networks (IJCNN)*, July 2014, pp. 29–34.
- [48] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

- [49] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.
- [50] M. Shahsavari, P. Falez, and P. Boulet, "Combining a volatile and nonvolatile memristor in artificial synapse to improve learning in spiking neural networks," in *2016 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, July 2016, pp. 67–72.
- [51] A. Polsky, B. Mel, and J. Schiller, "Encoding and decoding bursts by nmda spikes in basal dendrites of layer 5 pyramidal neurons," *Journal of Neuroscience*, vol. 29, no. 38, pp. 11 891–11 903, 2009.
- [52] M. Prezioso, F. Merrih-Bayat, B. Hoskins, G. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, no. 7550, p. 61, 2015.
- [53] M. Hu, Y. Chen, J. J. Yang, Y. Wang, and H. H. Li, "A compact memristor-based dynamic synapse for spiking neural networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 36, no. 8, pp. 1353–1366, Aug 2017.
- [54] P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang, W. Zhang, Q. Zhang, N. Deng, L. Shi, H.-S. P. Wong *et al.*, "Face classification using electronic synapses," *Nature communications*, vol. 8, p. 15199, 2017.
- [55] D. Soudry, D. D. Castro, A. Gal, A. Kolodny, and S. Kvatinsky, "Memristor-based multilayer neural networks with online gradient descent training," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2408–2421, Oct 2015.
- [56] H. Kim, M. P. Sah, C. Yang, T. Roska, and L. O. Chua, "Memristor bridge synapses," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 2061–2070, June 2012.
- [57] S. P. Adhikari, H. Kim, R. K. Budhathoki, C. Yang, and L. O. Chua, "A circuit-based learning architecture for multilayer neural networks with memristor bridge synapses," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 1, pp. 215–223, Jan 2015.
- [58] Y. S. Kim and K. S. Min, "Synaptic weighting circuits for cellular neural networks," in *2012 13th International Workshop on Cellular Nanoscale Networks and their Applications*, Aug 2012, pp. 1–6.
- [59] Y. Zhang, Y. Li, X. Wang, and E. G. Friedman, "Synaptic characteristics of ag/aginsbte/ta-based memristor for pattern recognition applications," *IEEE Transactions on Electron Devices*, vol. 64, no. 4, pp. 1806–1811, 2017.
- [60] Y. Zhang, X. Wang, and E. G. Friedman, "Memristor-based circuit design for multilayer neural networks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 2, pp. 677–686, 2018.
- [61] F. Alibart, E. Zamanidoost, and D. B. Strukov, "Pattern classification by memristive crossbar circuits using ex situ and in situ training," *Nature communications*, vol. 4, p. 2072, 2013.
- [62] R. Hasan and T. M. Taha, "Enabling back propagation training of memristor crossbar neuromorphic processors," in *Neural Networks (IJCNN), 2014 International Joint Conference on*. IEEE, 2014, pp. 21–28.
- [63] A. M. Sheri, H. Hwang, M. Jeon, and B. g. Lee, "Neuromorphic character recognition system with two pcmo memristors as a synapse," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 6, pp. 2933–2941, June 2014.
- [64] Z. Wang, S. Joshi, S. E. Savelev, H. Jiang, R. Midya, P. Lin, M. Hu, N. Ge, J. P. Strachan, Z. Li *et al.*, "Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing," *Nature materials*, vol. 16, no. 1, p. 101, 2017.
- [65] Z. Wang, S. Joshi, S. Savelev, W. Song, R. Midya, Y. Li, M. Rao, P. Yan, S. Asapu, Y. Zhuo *et al.*, "Fully memristive neural networks for pattern classification with unsupervised learning," *Nature Electronics*, vol. 1, no. 2, p. 137, 2018.
- [66] E. Rosenthal, S. Greshnikov, D. Soudry, and S. Kvatinsky, "A fully analog memristor-based neural network with online gradient training," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2016, pp. 1394–1397.
- [67] L. Danial, N. Wainstein, S. Kraus, and S. Kvatinsky, "Didactic: A data-intelligent digital-to-analog converter with a trainable integrated circuit using memristors," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2017.
- [68] J. Shamsi, K. Mohammadi, and S. B. Shokouhi, "A hardware architecture for columnar-organized memory based on cmos neuron and memristor crossbar arrays," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2018.
- [69] Y. Jiang, P. Huang, D. Zhu, Z. Zhou, R. Han, L. Liu, X. Liu, and J. Kang, "Design and hardware implementation of neuromorphic systems with rram synapses and threshold-controlled neurons for pattern recognition," *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp. 1–13, 2018.
- [70] A. Chowdhury, A. Ayman, S. Dey, M. Sarker, and A. I. Arka, "Simulations of threshold logic unit problems using memristor based synapses and cmos neuron," in *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, Dec 2017, pp. 1–4.
- [71] M. Hu, H. Li, Y. Chen, Q. Wu, G. S. Rose, and R. W. Linderman, "Memristor crossbar-based neuromorphic computing system: A case study," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 10, pp. 1864–1878, 2014.
- [72] O. Krestinskaya, K. N. Salama, and A. P. James, "Analog backpropagation learning circuits for memristive crossbar neural networks," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.
- [73] R. Naous, M. AlShedivat, E. Neftci, G. Cauwenberghs, and K. N. Salama, "Memristor-based neural networks: Synaptic versus neuronal stochasticity," *AIP Advances*, vol. 6, no. 11, p. 111304, 2016.
- [74] X. Wu, V. Saxena, and K. Zhu, "Homogeneous spiking neuromorphic system for real-world pattern recognition," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 5, no. 2, pp. 254–266, June 2015.
- [75] W. Xinyu, S. V, and Z. Kehan, "A cmos spiking neuron for dense memristor-synapse connectivity for brain-inspired computing," in *2015 International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–6.
- [76] I. E. Ebong and P. Mazumder, "Cmos and memristor-based neural network design for position detection," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 2050–2060, 2012.
- [77] A. Irmanova, O. Krestinskaya, and A. P. James, "Neuromorphic adaptive edge-preserving denoising filter," in *2017 IEEE International Conference on Rebooting Computing (ICRC)*, Nov 2017, pp. 1–6.
- [78] J. Zhang and X. Liao, "Synchronization and chaos in coupled memristor-based fitzhugh-nagumo circuits with memristor synapse," *AEU-International Journal of Electronics and Communications*, vol. 75, pp. 82–90, 2017.
- [79] H. Jiang, W. Zhu, F. Luo, K. Bai, C. Liu, X. Zhang, J. J. Yang, Q. Xia, Y. Chen, and Q. Wu, "Cyclical sensing integrate-and-fire circuit for memristor array based neuromorphic computing," in *Circuits and Systems (ISCAS), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 930–933.
- [80] J. Shamsi, A. Amirsoleimani, S. Mirzakuchaki, A. Ahmade, S. Alirezaee, and M. Ahmadi, "Hyperbolic tangent passive resistive-type neuron," in *Circuits and Systems (ISCAS), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 581–584.
- [81] G. Khodabandehloo, M. Mirhassani, and M. Ahmadi, "Analog implementation of a novel resistive-type sigmoidal neuron," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 4, pp. 750–754, April 2012.
- [82] S. P. Adhikari, C. Yang, H. Kim, and L. O. Chua, "Memristor bridge synapse-based neural network and its learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 9, pp. 1426–1435, Sept 2012.
- [83] M. Al-Shedivat, R. Naous, E. Neftci, G. Cauwenberghs, and K. N. Salama, "Inherently stochastic spiking neurons for probabilistic neural computation," in *Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on*. IEEE, 2015, pp. 356–359.
- [84] M. Al-Shedivat, R. Naous, G. Cauwenberghs, and K. N. Salama, "Memristors empower spiking neurons with stochasticity," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 5, no. 2, pp. 242–253, June 2015.
- [85] D. Querlioz, O. Bichler, and C. Gamrat, "Simulation of a memristor-based spiking neural network immune to device variations," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2011, pp. 1775–1781.
- [86] A. P. James, I. Fedorova, T. Ibrayev, and D. Kudithipudi, "Htm spatial pooler with memristor crossbar circuits for sparse biometric recognition," *IEEE Transactions on Biomedical Circuits and Systems*, vol. PP, no. 99, pp. 1–12, 2017.
- [87] D. Fan, M. Sharad, A. Sengupta, and K. Roy, "Hierarchical temporal memory based on spin-neurons and resistive memory for energy-efficient brain-inspired computing," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 9, pp. 1907–1919, 2016.
- [88] O. Krestinskaya and A. P. James, "Approximate probabilistic neural networks with gated threshold logic," *arXiv preprint arXiv:1808.00733*, 2018.

- [89] A. Serb, J. Bill, A. Khiat, R. Berdan, R. Legenstein, and T. Prodromakis, "Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses," *Nature communications*, vol. 7, p. 12611, 2016.
- [90] O. Krestinskaya and A. P. James, "Binary weighted memristive analog deep neural network for near-sensor edge processing," *arXiv preprint arXiv:1808.00737*, 2018.
- [91] D. Soudry, D. Di Castro, A. Gal, A. Kolodny, and S. Kvatinisky, "Hebbian learning rules with memristors," *Israel Institute of Technology: Haifa, Israel*, 2013.
- [92] R. Senthilkumar and R. K. Gnanamurthy, "A detailed survey on 2d and 3d still face and face video databases part i," in *Communications and Signal Processing (ICCSP), 2014 International Conference on*, April 2014, pp. 1405–1409.
- [93] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang *et al.*, "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks," *Nature Communications*, vol. 9, no. 1, p. 2385, 2018.
- [94] P. Wijesinghe, A. Ankit, A. Sengupta, and K. Roy, "An all-memristor deep spiking neural computing system: A step toward realizing the low-power stochastic brain," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 5, pp. 345–358, Oct 2018.
- [95] S. Duan, X. Hu, Z. Dong, L. Wang, and P. Mazumder, "Memristor-based cellular nonlinear/neural network: Design, analysis, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1202–1213, June 2015.
- [96] L. O. Chua and L. Yang, "Cellular neural networks: theory," *IEEE Transactions on Circuits and Systems*, vol. 35, no. 10, pp. 1257–1272, Oct 1988.
- [97] S. N. Truong, S. Shin, J. Song, H. S. Mo, F. Corinto, and K. S. Min, "Memristor-based cellular nanoscale networks: Theory, circuits, and applications," in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2015, pp. 1134–1137.
- [98] X. Hu, G. Feng, S. Duan, and L. Liu, "A memristive multilayer cellular neural network with applications to image processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 8, pp. 1889–1901, Aug 2017.
- [99] M. Di Marco, M. Forti, and L. Pancioni, "Memristor standard cellular neural networks computing in the flux-charge domain," *Neural Networks*, 2017.
- [100] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Computation*, vol. 29, no. 9, pp. 2352–2449, Sept 2017.
- [101] M. T. McCann, K. H. Jin, and M. Unser, "Convolutional neural networks for inverse problems in imaging: A review," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 85–95, Nov 2017.
- [102] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Transactions on Computational Imaging*, vol. 2, no. 2, pp. 109–122, June 2016.
- [103] L. Xia, T. Tang, W. Huangfu, M. Cheng, X. Yin, B. Li, Y. Wang, and H. Yang, "Switched by input: Power efficient structure for rram-based convolutional neural network," in *2016 53rd ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2016, pp. 1–6.
- [104] T. Tang, L. Xia, B. Li, Y. Wang, and H. Yang, "Binary convolutional neural network on rram," in *2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC)*, Jan 2017, pp. 782–787.
- [105] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramanian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, June 2016, pp. 14–26.
- [106] L. Ni, Z. Liu, W. Song, J. J. Yang, H. Yu, K. Wang, and Y. Wang, "An energy-efficient and high-throughput bitwise cnn on sneak-path-free digital rram crossbar," in *2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, July 2017, pp. 1–6.
- [107] L. Ni, Z. Liu, H. Yu, and R. V. Joshi, "An energy-efficient digital rram-crossbar-based cnn with bitwise parallelism," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 3, pp. 37–46, Dec 2017.
- [108] C. Yakopcic, R. Hasan, and T. M. Taha, "Memristor based neuromorphic circuit for ex-situ training of multi-layer neural network algorithms," in *2015 International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–7.
- [109] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves *et al.*, "Analogue signal and image processing with large memristor crossbars," *Nature Electronics*, vol. 1, no. 1, p. 52, 2018.
- [110] J. Wang, S. Hu, X. Zhan, Q. Yu, Z. Liu, T. P. Chen, Y. Yin, S. Hosaka, and Y. Liu, "Handwritten-digit recognition by hybrid convolutional neural network based on hfo 2 memristive spiking-neuron," *Scientific reports*, vol. 8, no. 1, p. 12546, 2018.
- [111] T. Serrano-Gotarredona, T. Prodromakis, and B. Linares-Barranco, "A proposal for hybrid memristor-cmos spiking neuromorphic learning systems," *IEEE Circuits and Systems Magazine*, vol. 13, no. 2, pp. 74–88, Secondquarter 2013.
- [112] H. Mostafa, C. Mayr, and G. Indiveri, "Beyond spike-timing dependent plasticity in memristor crossbar arrays," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2016, pp. 926–929.
- [113] E. Covi, S. Brivio, A. Serb, T. Prodromakis, M. Fanciulli, and S. Spiga, "Hfo2-based memristors for neuromorphic applications," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2016, pp. 393–396.
- [114] N. Panwar, B. Rajendran, and U. Ganguly, "Arbitrary spike time dependent plasticity (stdp) in memristor by analog waveform engineering," *IEEE Electron Device Letters*, vol. 38, no. 6, pp. 740–743, June 2017.
- [115] I. E. Ebgong and P. Mazumder, "Cmos and memristor-based neural network design for position detection," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 2050–2060, June 2012.
- [116] Z. Wang, M. Rao, J.-W. Han, J. Zhang, P. Lin, Y. Li, C. Li, W. Song, S. Asapu, R. Midya *et al.*, "Capacitive neural network with neuro-transistors," *Nature communications*, vol. 9, no. 1, p. 3208, 2018.
- [117] L. Deng, G. Li, N. Deng, D. Wang, Z. Zhang, W. He, H. Li, J. Pei, and L. Shi, "Complex learning in bio-plausible memristive networks," *Scientific reports*, vol. 5, p. 10684, 2015.
- [118] S.-W. Lee and H.-H. Song, "A new recurrent neural-network architecture for visual pattern recognition," *IEEE Transactions on Neural Networks*, vol. 8, no. 2, pp. 331–340, Mar 1997.
- [119] G. Bao, Z. Zeng, and Y. Shen, "Region stability analysis and tracking control of memristive recurrent neural network," *Neural Networks*, vol. 98, pp. 51–58, 2018.
- [120] H. Liu, Z. Wang, and B. Shen, "Discrete-time memristive recurrent neural networks with time-varying delays: Exponential stability analysis," in *2016 35th Chinese Control Conference (CCC)*, July 2016, pp. 3584–3589.
- [121] G. M. T. Xavier, F. G. Castañeda, L. M. F. Nava, and J. A. M. Cadenas, "Memristive recurrent neural network," *Neurocomputing*, vol. 273, pp. 281–295, 2018.
- [122] Y. Maeda and M. Wakamura, "Simultaneous perturbation learning rule for recurrent neural networks and its fpga implementation," *IEEE Transactions on Neural Networks*, vol. 16, no. 6, pp. 1664–1672, Nov 2005.
- [123] K. Smagulova, K. Adam, O. Krestinskaya, and A. P. James, "Design of cmos-memristor circuits for lstm architecture," *arXiv preprint arXiv:1806.02366*, 2018.
- [124] K. Adam, K. Smagulova, and A. P. James, "Memristive lstm network hardware architecture for time-series predictive modeling problem," *arXiv preprint arXiv:1809.03119*, 2018.
- [125] A. James, T. Ibrayev, O. Krestinskaya, and I. Dolzhikova, "Introduction to memristive htm circuits," in *Memristor and Memristive Neural Networks*. InTech, 2018.
- [126] A. Martinez and R. Benavente, "The ar face database," *Rapport technique*, vol. 24, 1998.
- [127] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.
- [128] A. Irmanova and A. P. James, "Neuron inspired data encoding memristive multi-level memory cell," *Analog Integrated Circuits and Signal Processing*, pp. 1–6, 2018.
- [129] O. Krestinskaya and A. P. James, "Feature extraction without learning in an analog spatial pooler memristive-cmos circuit design of hierarchical temporal memory," *Analog Integrated Circuits and Signal Processing*, pp. 1–9, 2018.
- [130] T. Ibrayev, U. Myrzakhan, O. Krestinskaya, A. Irmanova, and A. P. James, "On-chip face recognition system design with memristive hierarchical temporal memory," *arXiv preprint arXiv:1709.08184*, 2017.
- [131] A. James, T. Ibrayev, and O. Krestinskaya, "Design and implication of a rule based weight sparsity module in htm spatial pooler," in *Electronics, Circuits and Systems (ICECS), 2017 24th IEEE International*. IEEE, 2017.
- [132] S. Xiao, X. Xie, S. Wen, Z. Zeng, T. Huang, and J. Jiang, "Gst-memristor-based online learning neural networks," *Neurocomputing*, vol. 272, pp. 677–682, 2018.

- [133] N. Dastanova, S. Duisenbay, O. Krestinskaya, and A. P. James, "Bit-plane extracted moving-object detection using memristive crossbar-cam arrays for edge computing image devices," *IEEE Access*, vol. 6, pp. 18 954–18 966, 2018.
- [134] D. Mikhailenko, C. Liyanagedera, A. P. James, and K. Roy, "M 2 ca: Modular memristive crossbar arrays," in *Circuits and Systems (ISCAS), 2018 IEEE International Symposium on*. IEEE, 2018, pp. 1–5.
- [135] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future generation computer systems*, vol. 29, no. 7, pp. 1645–1660, 2013.
- [136] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1125–1142, Oct 2017.
- [137] K. Eshraghian, K. R. Cho, O. Kavehei, S. K. Kang, D. Abbott, and S. M. S. Kang, "Memristor mos content addressable memory (mcam): Hybrid architecture for future high performance search engines," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, no. 8, pp. 1407–1417, Aug 2011.
- [138] H. Jiang, L. Han, P. Lin, Z. Wang, M. H. Jang, Q. Wu, M. Barnell, J. J. Yang, H. L. Xin, and Q. Xia, "Sub-10 nm ta channel responsible for superior performance of a hfo 2 memristor," *Scientific reports*, vol. 6, p. 28525, 2016.
- [139] I. Vourkas, D. Stathis, G. C. Sirakoulis, and S. Hamdioui, "Alternative architectures toward reliable memristive crossbar memories," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 1, pp. 206–217, Jan 2016.
- [140] L. Chen, Z.-Y. He, T.-Y. Wang, Y.-W. Dai, H. Zhu, Q.-Q. Sun, and D. W. Zhang, "Cmos compatible bio-realistic implementation with ag/hfo2-based synaptic nanoelectronics for artificial neuromorphic system," *Electronics*, vol. 7, no. 6, p. 80, 2018.
- [141] S. Kim, H. Kim, S. Hwang, M.-H. Kim, Y.-F. Chang, and B.-G. Park, "Analog synaptic behavior of a silicon nitride memristor," *ACS applied materials & interfaces*, vol. 9, no. 46, pp. 40 420–40 427, 2017.
- [142] T. Cabaret, L. Fillaud, B. Jousselme, J. O. Klein, and V. Derycke, "Electro-grafted organic memristors: Properties and prospects for artificial neural networks based on stdp," in *14th IEEE International Conference on Nanotechnology*, Aug 2014, pp. 499–504.
- [143] Y.-F. Chang, B. Fowler, Y.-C. Chen, F. Zhou, C.-H. Pan, T.-C. Chang, and J. C. Lee, "Demonstration of synaptic behaviors and resistive switching characterizations by proton exchange reactions in silicon oxide," *Scientific reports*, vol. 6, p. 21268, 2016.
- [144] F. Alibart, L. Gao, B. D. Hoskins, and D. B. Strukov, "High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm," *Nanotechnology*, vol. 23, no. 7, p. 075201, 2012.
- [145] R. Naous, M. Al-Shedivat, and K. N. Salama, "Stochasticity modeling in memristors," *IEEE Transactions on Nanotechnology*, vol. 15, no. 1, pp. 15–28, 2016.
- [146] J. J. Yang, M.-X. Zhang, J. P. Strachan, F. Miao, M. D. Pickett, R. D. Kelley, G. Medeiros-Ribeiro, and R. S. Williams, "High switching endurance in tao x memristive devices," *Applied Physics Letters*, vol. 97, no. 23, p. 232102, 2010.
- [147] E. Amat, A. Rubio *et al.*, "Memristive crossbar memory lifetime evaluation and reconfiguration strategies," *IEEE Transactions on Emerging Topics in Computing*, vol. 6, no. 2, pp. 207–218, 2018.
- [148] A. Fantini, L. Goux, R. Degraeve, D. Wouters, N. Raghavan, G. Kar, A. Belmonte, Y.-Y. Chen, B. Govoreanu, and M. Jurczak, "Intrinsic switching variability in hfo 2 rram," in *Memory Workshop (IMW), 2013 5th IEEE International*. IEEE, 2013, pp. 30–33.
- [149] K. M. Kim, J. J. Yang, J. P. Strachan, E. M. Grafals, N. Ge, N. D. Melendez, Z. Li, and R. S. Williams, "Voltage divider effect for the improvement of variability and endurance of tao x memristor," *Scientific reports*, vol. 6, p. 20085, 2016.
- [150] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano letters*, vol. 10, no. 4, pp. 1297–1301, 2010.
- [151] K. Inc., "Knowm memristors," Tech. Rep., 2015. [Online]. Available: https://knowm.org/downloads/Knowm_Memristors.pdf
- [152] L. Xie, H. A. Du Nguyen, M. Taouil, S. Hamdioui, and K. Bertels, "Interconnect networks for memristor crossbar," in *Nanoscale Architectures (NANOARCH), 2015 IEEE/ACM International Symposium on*. IEEE, 2015, pp. 124–129.
- [153] H. A. Du Nguyen, L. Xie, J. Yu, M. Taouil, and S. Hamdioui, "Interconnect networks for resistive computing architectures," in *2017 12th International Conference on Design & Technology of Integrated Systems In Nanoscale Era (DTIS)*. IEEE, 2017, pp. 1–6.
- [154] P. Mazumder, S.-M. Kang, and R. Waser, "Memristors: devices, models, and applications," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1911–1919, 2012.
- [155] F. M. Bayat, B. Hoskins, and D. B. Strukov, "Phenomenological modeling of memristive devices," *Applied Physics A*, vol. 118, no. 3, pp. 779–786, 2015.
- [156] D. Biolek, Z. Kolka, V. Biolková, Z. Biolek, M. Potřebič, and D. Tošić, "Modeling and simulation of large memristive networks," *International Journal of Circuit Theory and Applications*, vol. 46, no. 1, pp. 50–65, 2018.
- [157] J. Singh and B. Raj, "Comparative analysis of memristor models and memories design," 2018.
- [158] M. D. Pickett, D. B. Strukov, J. L. Borghetti, J. J. Yang, G. S. Snider, D. R. Stewart, and R. S. Williams, "Switching dynamics in titanium dioxide memristive devices," *Journal of Applied Physics*, vol. 106, no. 7, p. 074508, 2009.
- [159] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *nature*, vol. 453, no. 7191, p. 80, 2008.
- [160] T. Prodromakis, B. P. Peh, C. Papavassiliou, and C. Toumazou, "A versatile memristor model with nonlinear dopant kinetics," *IEEE transactions on electron devices*, vol. 58, no. 9, pp. 3099–3105, 2011.
- [161] Y. N. Joglekar and S. J. Wolf, "The elusive memristor: properties of basic electrical circuits," *European Journal of Physics*, vol. 30, no. 4, p. 661, 2009.
- [162] Z. Biolek, D. Biolek, and V. Biolkova, "Spice model of memristor with nonlinear dopant drift," *Radioengineering*, vol. 18, no. 2, 2009.
- [163] S. Kvatinsky, E. G. Friedman, A. Kolodny, and U. C. Weiser, "Team: Threshold adaptive memristor model," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 1, pp. 211–221, Jan 2013.
- [164] D. Biolek, Z. Kolka, V. Biolkova, and Z. Biolek, "Memristor models for spice simulation of extremely large memristive networks," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2016, pp. 389–392.
- [165] I. Messaris, A. Serb, S. Stathopoulos, A. Khiat, S. Nikolaidis, and T. Prodromakis, "A data-driven verilog-a reram model," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2018.
- [166] L. Chua, "Memristor—the missing circuit element," *IEEE Transactions on Circuit Theory*, vol. 18, no. 5, pp. 507–519, Sep 1971.
- [167] D. Batas and H. Fiedler, "A memristor spice implementation and a new approach for magnetic flux-controlled memristor modeling," *IEEE Transactions on Nanotechnology*, vol. 10, no. 2, pp. 250–255, 2011.
- [168] . Rak and G. Cserey, "Macromodeling of the memristor in spice," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 4, pp. 632–636, April 2010.
- [169] S. Benderli and T. Wey, "On spice macromodelling of tio 2 memristors," *Electronics letters*, vol. 45, no. 7, pp. 377–379, 2009.
- [170] V. Keshmiri, "A study of the memristor models and applications," 2014.
- [171] H. Kim, M. P. Sah, C. Yang, T. Roska, and L. O. Chua, "Neural synaptic weighting with a pulse-based memristor circuit," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, no. 1, pp. 148–158, Jan 2012.
- [172] J. P. Strachan, A. C. Torrezan, F. Miao, M. D. Pickett, J. J. Yang, W. Yi, G. Medeiros-Ribeiro, and R. S. Williams, "State dynamics and modeling of tantalum oxide memristors," *IEEE Transactions on Electron Devices*, vol. 60, no. 7, pp. 2194–2202, 2013.
- [173] H. Abdalla and M. D. Pickett, "Spice modeling of memristors," in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, May 2011, pp. 1832–1835.
- [174] J. G. Simmons, "Generalized formula for the electric tunnel effect between similar electrodes separated by a thin insulating film," *Journal of applied physics*, vol. 34, no. 6, pp. 1793–1803, 1963.