

Adaptive Synaptogenesis Implemented on a Nanomagnetic Platform

Faiyaz Elahi Mullick,^{1,*} Supriyo Bandyopadhyay,^{2,†} Rob Baxter,^{3,‡} Tony J. Ragucci,^{4,§} and Avik W. Ghosh^{1,5,¶}

¹*Department of Electrical and Computer Engineering, University of Virginia*

²*Department of Electrical and Computer Engineering, Virginia Commonwealth University*

³*Department of Neurosurgery, University of Virginia*

⁴*Leonardo DRS*

⁵*Department of Physics, University of Virginia*

The human brain functions very differently from artificial neural networks (ANN) and possesses unique features that are absent in ANN. An important one among them is “adaptive synaptogenesis” that modifies synaptic weights when needed to avoid *catastrophic forgetting* and promote lifelong learning. The key aspect of this algorithm is supervised Hebbian learning, where weight modifications in the neocortex driven by temporal coincidence are further accepted or vetoed by an added control mechanism from the hippocampus during the training cycle, to make distant synaptic connections highly sparse and strategic. In this work, we discuss various algorithmic aspects of adaptive synaptogenesis tailored to edge computing, demonstrate its function using simulations, and design nanomagnetic hardware accelerators for specific functions of synaptogenesis.

I. INTRODUCTION

The need for brain-inspired computing far beyond conventional artificial neural networks (ANN) is a pressing one. Conventional machine learning (ML) based on Deep Neural Nets (DNN) is ill-suited for the rapid growth in edge intelligence owing to the considerable cost of wiring, exorbitant energy demands, and a limit on memory resources. In ANN/DNN, new sensory data are learned independent of past history, requiring added hardware and costly synaptic interconnects, with an exploding area footprint and energy budget. In edge environments requiring on-chip computing, such as autonomous robots exploring mines, battlefield vehicles navigating enemy terrain, underwater drones or Mars rovers, robots must analyze rapidly varying situations with very limited memory resources and energy budgets, without relying on a cloud that is either unavailable or unreliable in the face of security breaches. In 3D Autonomous Simultaneous Localization and Mapping (ASLAM) for mobile robotics, for instance, robot actions need to be calculated on the go offline, but this is traditionally done by deleting out-of-date data using a delayed nearest neighbor data association strategy [1–3]. Going beyond such a local adaptive mapping and navigation, one step at a time, becomes completely prohibitive over an expanded time horizon.

In a distributed neural net, the challenge with learning sequential data is called the *stability-plasticity dilemma*, the choice between integrating new knowledge vs remembering previously acquired knowledge. The conventional way to deal with this is to reduce overlap among

the stored internal representations, for example by using sparse or interleaved learning. Connection networks must then absorb new inputs and adjust synaptic weights by *small increments*, thereby preventing sequential learning. Attempts to address this challenge have mostly been made in software, for instance, a Fahlman offset [4] in the derivative of the sigmoid function in backpropagation to avoid entrenchment in its flat parts. Intel’s recent AI robot [5] relied on a learning phase where prototype data are moved around but not erased in feature space, punishing the wrong category or rewarding the right category, and only allocating new resources if the error persists and an unknown category is thereby identified. Extending learning to a practical hardware environment, with acceptable Size Weight and Power (SWaP) is still an unrealized target. In contrast, human cognitive abilities provide a remarkable example of how the brain encodes sequential learning very efficiently with a fixed memory bank, avoiding **catastrophic forgetfulness** (sequential data over-write) with incremental and contextual learning. The evolved trick is an over-riding hippocampus (HC) that acts as a memory support structure for a severely resource-constrained neocortex (NC), with dense local but very sparse, one in a million, non-local connectivity that mediates perceptions and decisions. In contrast to brain development which proceeds from sensory, up through the cortical hierarchy (i.e., bottom-up), later learning proceeds in a top-down manner directed by the hippocampal system, a dense recurrently connected generative network with the ability to process and correlate long term multi-dimensional sequences [6, 7] to build a sophisticated “World Model” for the organism. In addition to the hardware support structure of a dual memory, the brain also uses algorithmic techniques to encode episodic memory by hierarchically implementing depths of representation. Big ideas are built out of association between contexts, event histories and multi-sensory microscopic details; for instance reframing a ‘zebra’ as a

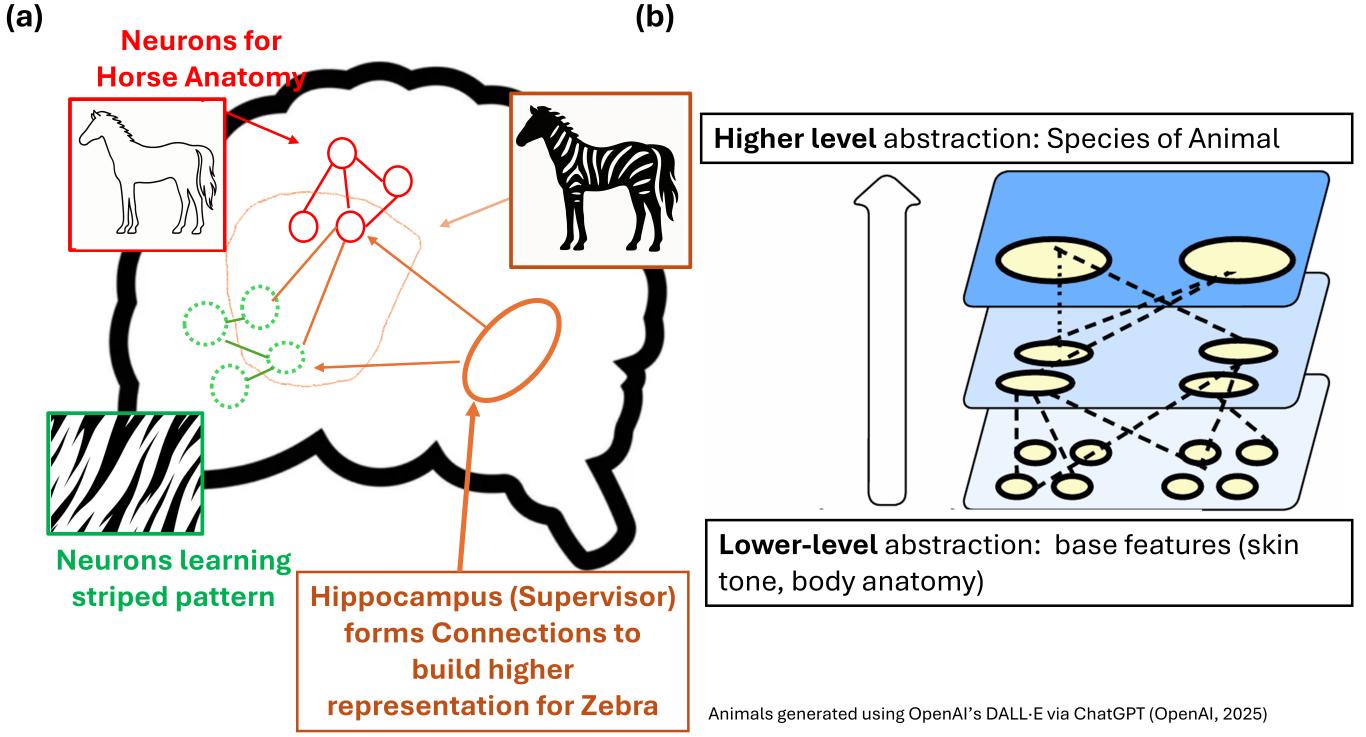
* fm4fv@virginia.edu

† sbandy@vcu.edu

‡ rbaxter37@gmail.com

§ tony.ragucci@drs.com

¶ ag7rq@virginia.edu



Animals generated using OpenAI's DALL-E via ChatGPT (OpenAI, 2025)

Figure 1: (a) lower level representations of stripes and animal anatomy learnt by groups of neurons that will get wired together to (b) represent a higher level abstraction such as an animal species of zebra

'striped horse'.(Fig. 1).

*The process of having a separate dense hippocampal network directing a sparser neocortical network allows us to avoid catastrophic forgetfulness and compose coherent long-term semantic memory (**lifelong learning**) from short, fragmented, conditionally independent, episodic events and their intermediate representations.*

II. BRAIN INSPIRED LEARNING MODELS

The goal of this section is to develop a mathematical model for the hippocampal-neocortical (HC-NC) interaction that allows a compute engine to maintain sparsity while encoding episodic associative memory - i.e., memory encoded as a sequence of time-correlated (but not necessarily causal) events. The key part of this algorithm is the employment of a digital synaptogenesis step - namely, the creation or removal (one or zero) of primarily distal synaptic connections, on top of an analog weight modification step, where the synaptic weights in the NC are increased when they co-fire in a Hebbian format ('neurons that fire together wire together'). In contrast to conventional Hebbian networks which are unsupervised, in this model the Hebbian coefficient is a probabilistic random variable, and on top of that the HC exercises an added supervisory control through digital

synaptogenesis. In other words, this process is *partially supervised*.

We will introduce the mathematical algorithms in this section. In section III, we will discuss the implications of these algorithms with a model problem, while in section IV we will explore a potential hardware framework to realize this architecture on-chip using compact voltage tunable magnetic neurons.

A. Hebbian Learning

The most widespread synaptic modification rule for biological models of the brain employ Hebbian learning, where weights are updated based on the time proximity of neuronal firing (Fig. 2). Let us start with simple Hebbian neural network H , and define the main variables. Let the input vector be $X = \{x_1, x_2, x_3, \dots, x_n\}$ where x_i can be any analog value between 0 and 1. We do this because conventional datasets are continuous such as temperature readings from a factory sensor, pixel intensity in a 2D image, current values from a photo-detector etc. and are generally normalized between 0 and 1. Alongside this, we also have a set of dimensionless weights $W = \{w_{11}, w_{12}, \dots, w_{mn}\}$ where w_{ij} is once again a continuous value between 0 and 1. For each input vector \mathbf{X} , an output vector $Y = \{y_1, y_2, y_3, \dots, y_m\}$ exists in the

training dataset where the definition of \mathbf{Y} depends on the nature of the task:

- Regression : Since the objective is to predict continuous values, each y_j represents a real-valued prediction corresponding to a continuous variable and m represents how many outputs the model will learn to predict. As such, the outputs are typically normalized to fall within 0 and 1. Formally, for each input \mathbf{X} , the output is:

$$Y = H(X; W) \quad (1)$$

Where, the Hebbian model \mathbf{H} will predict \mathbf{Y} using \mathbf{W} and \mathbf{X} .

- Classification : Here the output represents a probability distribution over m discrete classes and each y_j indicates the predicted probability that the input \mathbf{X} belongs to class j and the vector satisfies the constraint:

$$\sum_{j=1}^m y_j = 1 \quad (2)$$

This ensures that the output forms a valid probability distribution. The predicted class is typically chosen as the index with the highest probability (magnitude):

$$j_{max} = \arg \max_j y_j \quad (3)$$

The ground truth label (original labels from the dataset) are represented using one-hot encoding such that the predicted label \mathbf{Y}^* will have elements:

$$y_j^* = \begin{cases} 1, & \text{if } j = j_{max} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Let us provide a concrete example. Let us say you have a 2D image dataset of cats, dogs and deer. Your dataset input vector size will be determined by the number of pixels each image has, but the size of your output vector will be 3 since there are three labels. The one hot encoded label for each would be 100, 010 and 001 respectively. This means $Y = \{y_1, y_2, y_3\}$ where y_j follows (2). In actual practice, the predicted output y_j^* 's will not exactly match the one hot encoded labels. Instead, it will be a distribution. For example, if the input image was that of a cat whose label is 100, the output \mathbf{Y} might look like $y_j = \{0.8, 0.15, 0.05\}$ where the maximum of 0.8 indicates the model predicted the image is a cat with 80 % confidence.

Now, a Hebbian model does not use a loss function L like a traditional network and instead the extent to which we will change any weight w_{ij} depends on the time difference Δt between an input x_i arriving and the excitation

of any output y_j (Fig. 2(b)). This makes our Hebbian weight update rule at any time (t) as:

$$\Delta w_{ij}(t) = \begin{cases} x_i(t) \cdot y_j(t) \cdot A_+ \cdot e^{-\Delta t/\tau_+}, & \text{if } \Delta t > 0 \\ x_i(t) \cdot y_j(t) \cdot A_- \cdot e^{\Delta t/\tau_-}, & \text{if } \Delta t < 0 \end{cases} \quad (5)$$

where A_+ and A_- are scaling factors (learning rate), Δt is the time between an input x_i arriving and an output y_j firing, and τ_+, τ_- are time constants for potentiation and depression, respectively. This continuous-time model captures the idea that causally aligned spikes (input before output) lead to synaptic strengthening, while anti-causal spikes (input after output) lead to weakening.

For our computation, we discretize time into n discrete timesteps t_n such that inputs and outputs are sampled in intervals of δt obeying $t_n = n \cdot \delta t$ where $n \in \mathbb{N}$. We also simplify (5) by assuming both x_i and y_j must be non-zero at the same t_n (i.e., replace the exponents with a ‘box’ function, and focus here on potentiation alone). This means we assume $\Delta t \rightarrow 0$ i.e. weight update will only happen at t_n :

$$\Delta w_{ij}(t_n) = \begin{cases} \epsilon \cdot x_i(t_n) \cdot y_j(t_n), & \text{if } x_i(t_n), y_j(t_n) > 0 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where ϵ combines both A_+, A_- into a single learning rate, and $\Delta W_{ij}(t_n)$ only updates at timestep t_n when both input and output are non-zero at t_n . We further modify (6) by adding an upper bound

$$\Delta w_{ij}(t_n) = \epsilon \left(x_i(t_n) - w_{ij}(t_n) - E(x) \right) y_j(t_n), \quad (7)$$

which ensures that the weights decay to zero when the feature is learned completely and fluctuations of the input x_i around the expectation $E(x)$ minimize. Also, the weights will decrease in magnitude anytime an output fires in the absence of an input i.e. a ‘false’ positive. We can then find the excitation as :

$$y_j = f \left(\frac{\mathbf{W}_j^T \mathbf{X}}{\sum_i^n w_{ij} + A} \right), \quad (8)$$

where we divide by the sum of lateral weights plus a constant A between 0 and 1 to constrain the excitation. We use the factor A for further control on this inhibition since initially the lateral sum $\sum_i^n w_{ij}$ might not be large enough to control the output.

We then pass the output y_j through an activation function f where f can be [8] a perceptron, sigmoid, tanh, ReLU, GeLU etc. and the nature of y_j will depend on the function being used. For example, with tanh, y_j is in the range of -1 and $+1$ while with ReLU $y_j = \frac{\mathbf{W}_j^T \mathbf{X}}{\sum_i^n w_{ij} + A}$ if $y_j > \theta_y$ (a given threshold on the output), otherwise 0.

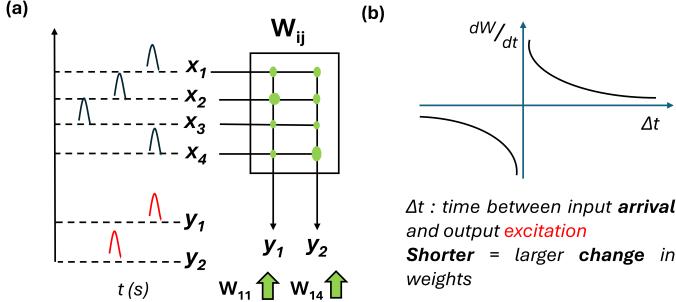


Figure 2: (a) Hebbian learning principle, y_1 firing just as inputs arrive at x_1 and x_4 implies weights W_{11} and W_{14} must be updated since Hebbian employs (b) spike timing dependent plasticity where a smaller Δt causes larger change ΔW .

B. Synaptogenesis and Shedding

Whenever new information is presented to biological brains, they either modify existing neural connections or form new ones. While classical artificial neural networks also modify connections, where they depart from biology is the formation of new connections. In ANN, all neurons are pre-connected and their weights are imposed during a training step, while in biological networks, many (most) synaptic connections are not even eligible. This biological phenomenon is called *synaptogenesis*. By using a combination of Hebbian weight modification and synaptogenesis, biological brains learn and remember new information.

One major modeling question for synaptogenesis is what guides the formation of new connections? While many complex methods exist, in our approach [9] we chose to use the average firing rate of a neuron to guide synapse formation. We define the calculation of the average firing rate using the following equation:

$$\bar{y}_j(t_n) = (1 - \alpha)\bar{y}_j(t_n - 1) + \alpha y_j(t_n - 1) \quad (9)$$

where the current average firing rate $\bar{y}_j(t)$ is a running average calculated using the previous average $\bar{y}_j(t_n - 1)$ and a ‘portion’ of the current instantaneous (no bar) output $y_j(t_n - 1)$ at each timestep (t_n). The variable α controls how much of the current output is used to update this running average. The key idea is that, for any output neuron $y_j = \sum_{i=1}^n w_{ij} x_i$ connected to n input neurons x_i via synapses w_{ij} (Fig. 3a), every time inputs excite x_i , they should end up firing y_j . We expect y_j to fire consistently with the frequency of inputs presented to x_i . However, sometimes the inputs might not be able to generate a strong enough excitation on the output. This is when synaptogenesis occurs as shown in Fig. 3(a) where new synapses are added if the average firing rate falls below some threshold ρ . Synapses are formed followed by modification using (7).

Once synapses are added and modified, a final scan is

done to check for any weights that are contributing too little to the excitation to incorporate a synaptic shedding step to reinforce sparsity. This is as simple as having a shedding threshold, so that any weights falling below are removed from W .

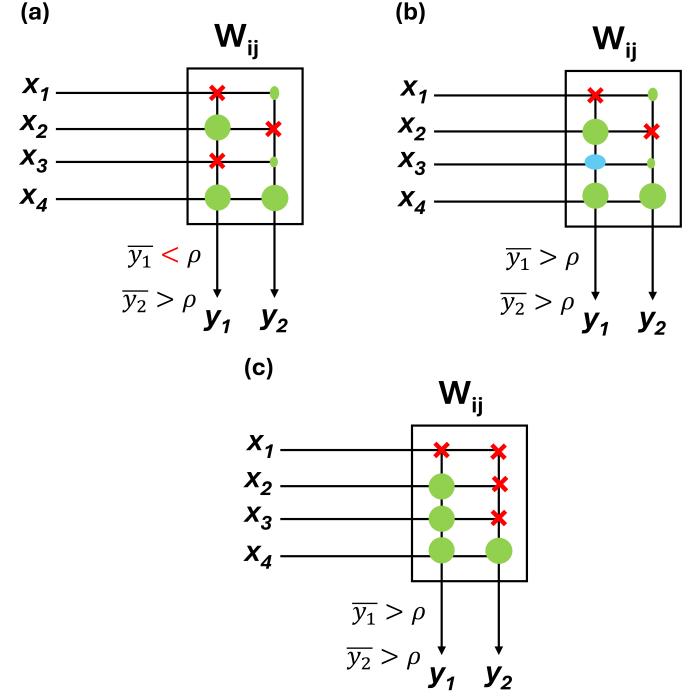


Figure 3: (a) Initial state where inputs are not stimulating y_1 enough, so its firing rate is low. (b) Synaptogenesis forms a new connection, increasing the firing rate above the threshold. Also, notice the smaller weights connected to y_2 . (c) Low-value weights are ‘shed,’ and the firing rate remains unaffected, indicating they were not contributing to the activation of y_2 .

Biological brains do not simply form connections immediately after a neuron fails to excite properly, they also consider the cost of such an action. They maintain sparsity and efficiency by only forming connections when a neuron is constantly excited and fails to activate. This means that there is a probabilistic aspect to synaptogenesis – the probability of the brain forming new connection(s) is less (often much less) than unity. To incorporate this into Hebbian learning, we first introduce an additional matrix of 0-s and 1-s called the *connection matrix* C_{ij} . This matrix tracks which inputs x_i are connected to each output y_j . This converts equations (7), (8) into:

- The modified weight update rule:

$$\Delta w_{ij} = \epsilon C_{ij} \left(x_i - w_{ij} - E(x) \right) y_j \quad (10)$$

- The new output excitation equation:

$$y_j = f \left(\frac{\mathbf{C}_j^T \mathbf{W}_j^T \mathbf{X}}{\sum_i^n w_{ij} + A} \right) \quad (11)$$

The connection matrix C_{ij} responsible for tracking synaptogenesis ensures that only connected synapses update their weights.

We now codify probabilistic synaptogenesis in a set of equations that control the elements of C_{ij} based on the average firing rate of each output neuron.

- For any input x_i exciting outputs y_j , we find the running average firing rates of y_j using (9).
- For each connection between x_i 's and y_j 's, we find the associated probabilities p_{ij} as

$$p_{ij} = \gamma(1 - C_{ij})a_i \quad (12)$$

where γ is the synapse formation rate and lies between 0 and 1. A value of 1 would imply fully deterministic synapse formation and a value of 0.02 would imply roughly 2% connectivity. We use a variable a_i to check for the presence of input x_i , i.e., $a_i = 1$ for $x_i > 0$ and 0 otherwise. This allows the probability to only be computed when an input is present i.e. non-zero.

- We update the connection matrix according to the following rule. For each p_{ij} sampled from a uniform distribution $B \sim U(0, 1)$ we change C_{ij} thus:
 - if $p_{ij} > B_{ij}$, $C_{ij} = 1$ and $w_{ij} = W_o$, where W_o is the initial weight value for any new connection (this is chosen based on dataset distribution)
 - if $p_{ij} < B_{ij}$, then C_{ij} remains unchanged.

The random variable \mathbf{B} allows synaptogenesis to be probabilistic and γ controls how frequently synaptogenesis can occur.

Finally shedding occurs near the end:

- if any $W_{ij} < W_{shed}$, $C_{ij} = 0$, where W_{shed} is the threshold of cutting off redundant weights.

To recap, W is an analog weight that attempts to update using coincidence, while C is an overriding digital enabling circuit that enables or vetoes these attempted modifications. In biological systems, these operate on different time-scales or epochs. In our model, we assume they are evaluated concurrently at each step.

The combined set of rules comprehensively describes how *synaptogenesis* occurs.

III. MODEL SIMULATIONS

A. Demonstration of Synaptogenesis

To demonstrate how the model creates new connections, we set up a simple experiment. Let us use a classification example. At first, (Fig. 4(a)) we generate a synthetic distribution with three inputs x_1 , x_2 and x_3 (500 data points) for a model with two outputs y_1 , y_2 which can be either 0 or 1 i.e. a classification scenario. We are simply trying to define a dataset that has only two possible labels for three different inputs. For our activation function we will use the perceptron which makes $y_j = 1$ if $y_j > \theta_y$. We initialize the run with $\gamma = 1$, $\epsilon = 0.05$, $\alpha = 0.01$ and activation threshold $\theta_y = 0.02$ on y_j and train the model for a total of 60 epochs, where each epoch is defined as the time required to train the model over the entire dataset once. Whenever a new epoch is started, the input dataset is randomized and the model is trained again. For 500 data points we would be training the model for a total of 30,000 time steps, while monitoring three randomly chosen weight values. In the middle of the run, we purposefully prune one weight and notice the average firing rate of one of the outputs to fall (Fig. 4(d)) which causes the model to form a connection immediately to recover the firing rate.

Now that we have explained synaptogenesis in a toy model, let us model a higher dimensional data set to see how we can generate sparse connections in the weight matrix. We generate a synthetic data set of 1-D orthogonal vectors (Fig. 5(a)), where each unique orthogonal vector belongs to a distinct class (Fig. 5(b)). We run the model according to the parameters in Table I. Here we see how initially the connection count climbs up until the model finally sheds to an optimal number of connections in Fig. 5(c), resulting in a sparse final weight matrix (Fig. 5(d)).

Name	Value
synapse formation rate γ	0.001
moving average window α	0.001
learning rate ϵ	0.05
initial weight W_o	0.2
shed threshold W_{shed}	0.01
epochs	70
average firing rate ρ	0.1
inhibition factor A	0.001
input neurons x_i	80
output neurons y_j	100

Table I: Unsupervised Model Parameters

B. Supervised Synaptogenesis

While an unsupervised approach allows us to see how the model is functioning on a fundamental level, it is dif-

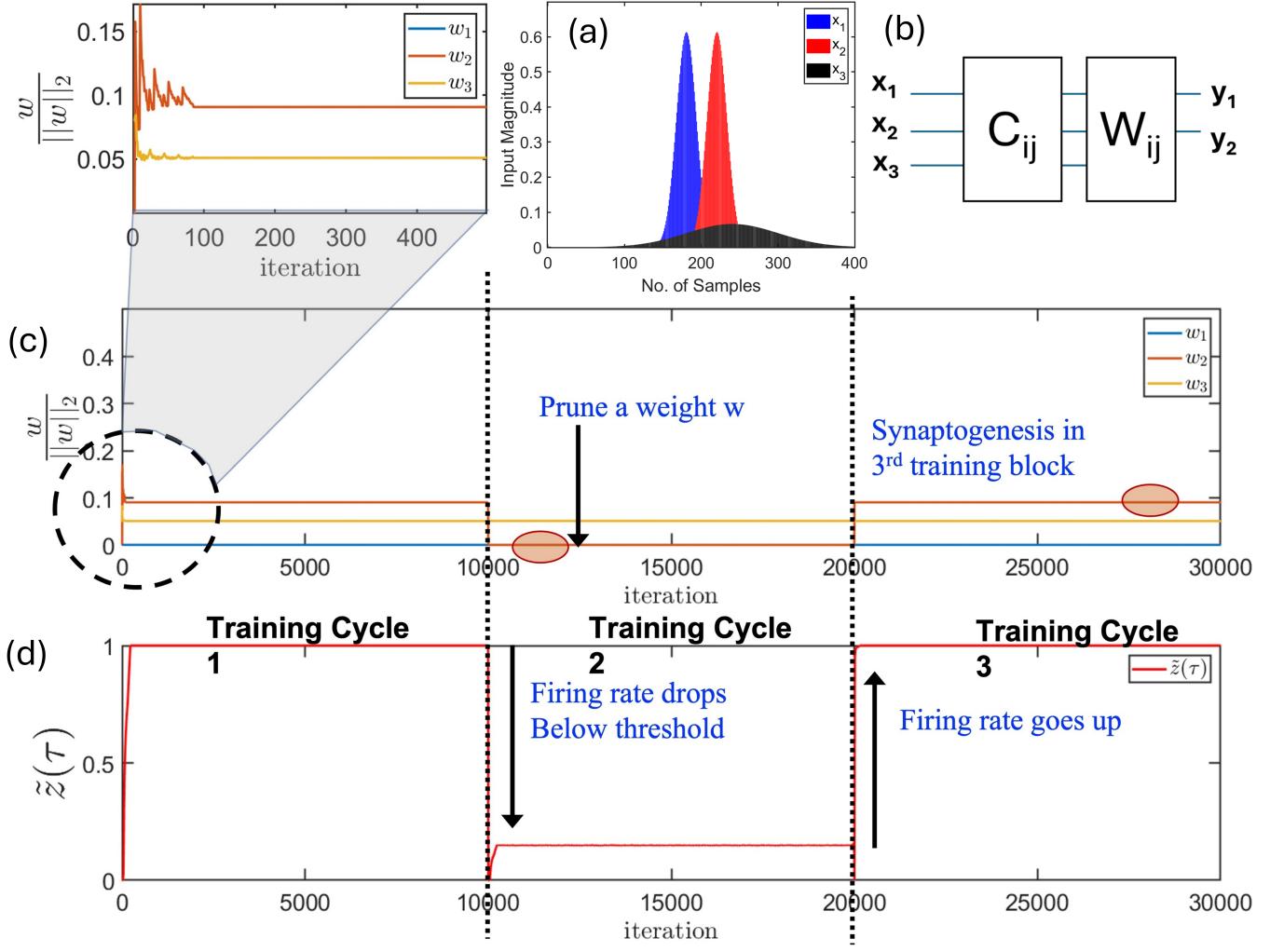


Figure 4: (a) Distribution of input dataset (500 points) run on (b) model with two outputs. (c) We monitor a few randomly selected weights over iterations (each iteration sends one datapoint across 60 epochs). (d) Pruned weight reduces firing rate which causes model to add weights until firing rate meets threshold

ficult to implement this in most applications, specially when we have labeled data sets. This is because not all applications benefit from data driven analysis, some examples such as image classification for labeled data **require** some sort of supervision to reward/punish the model to learn properly. For this reason, we modify the synaptic rules so that a supervision signal [9] guides the formation of synapses to maximize training accuracy. This expands the model into two layers (Fig. 6(b)). We now run a simulation to show that in a supervised scene, we see neuron reuse. We set up a scenario where we start out with 5 categories like in Fig. 5(a),(b) with 5 fully orthogonal unique vectors. We then generate variants with partial overlaps in each class shown in Fig. 6(a). For our simulation, we:

- Generate variants with partial overlaps in each class ensuring no interclass overlaps.

- Generate a total of 50 variants per class (overlap range 5% to 40%)
- Train a model with 0 variants, then 1 variant per class upto 50 variants per class. This means we have a total of 100 different models.
- Repeat training 50 times with different random seeds (for probabilistic synaptogenesis) for a total of 2500 models
- Analyze neuron count plots and test accuracy (80:20 split).
- compare accuracy to k -nearest neighbors for benchmarking

Fig. 7 shows the results of our runs. We see an initial rise in neuron count with increase in variation. However this

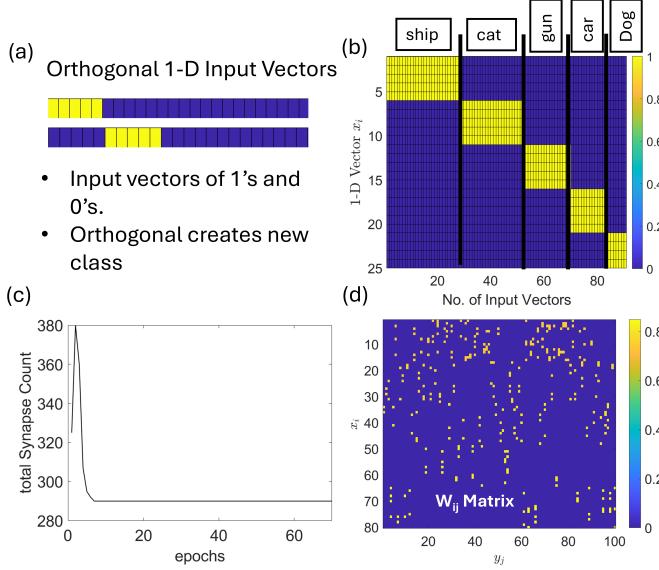


Figure 5: (a) 1-D generated binary vectors where orthogonality ensures distinct classes. (b) Dataset used with 5 classes, variation in no. of datavectors (c) Model optimizes number of synapses after a few epochs (d) Sparse W_{ij} matrix with higher number of points near the top since the associated inputs were excited the most.

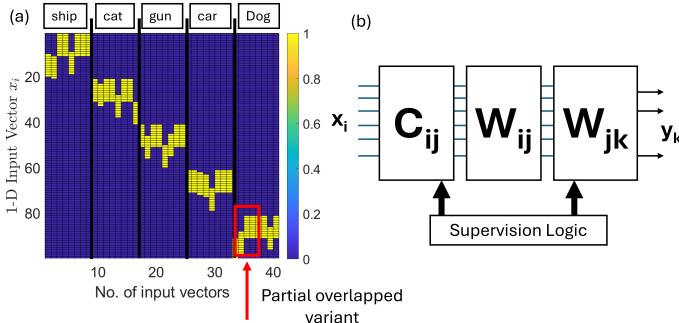


Figure 6: (a) Dataset with shared features (b) Supervised model

tapers off since the model no longer needs new neurons to reach maximum training accuracy. This shows that the dataset's shared features were learnt by the model and the model was able to re-use the same neurons. Accuracy is on par with k -nearest neighbors (k NN) for benchmarking reasons. We used sk-learn's packages for k NN.

IV. HARDWARE IMPLEMENTATION OF ADAPTIVE SYNAPTOGENESIS

We now discuss a set of hardware accelerators for adaptive synaptogenesis mainly comprising *nanomagnetic*

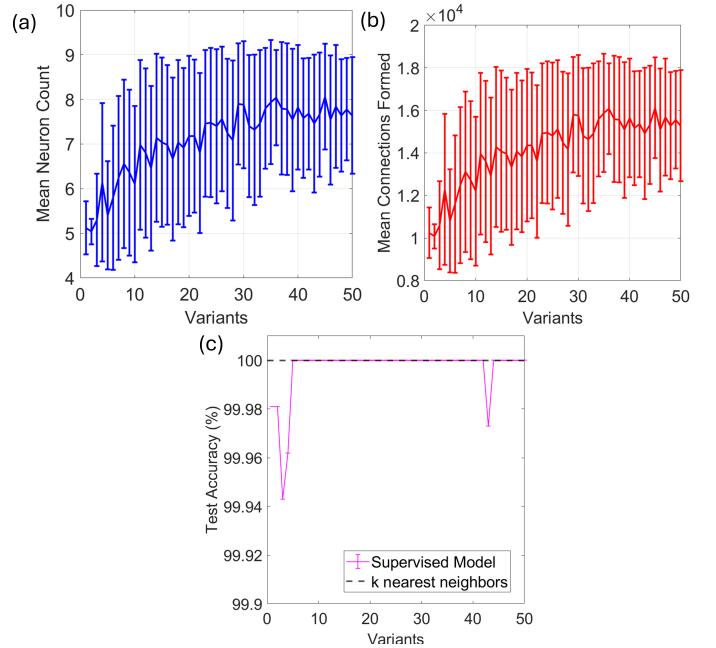


Figure 7: Average number of allocated (a) neurons (b) synapse connections employed during training of the supervised model using datasets with overlapping (shared features) and (c) their corresponding test set accuracy benchmarked against k nearest neighbors. We see neuron allocation plateau's as the shared features between input vectors allows re-use of neurons

Name	Value
synapse formation rate γ	1
moving average window α	0.001
learning rate ϵ	0.05
initial weight W_o	1
shed threshold W_{shed}	0.01
epochs	50
average firing rate ρ	0.1
inhibition factor A	1
Number of trials	50
input neurons x_i	1
hidden neurons y_j	50
Output neurons y_k	5
Output activation threshold θ_y	0.02

Table II: Supervised Model Parameters

netic devices for their superior energy-efficiency and non-volatility.

The principal equation for weight modification in adaptive synaptogenesis is Equation (7). It is clear from this equation that a hardware platform for adaptive synaptogenesis will require seven major components: (1) a device to measure the firing rate of a neuron, (2) a comparator for comparing the measured firing rate with a threshold, (3) an analog multiplier to multiply the analog (voltage or

current) outputs of two neurons, (4) an analog subtractor to subtract one neuronal output from another, (5) analog (preferably non-volatile) weights, (6) a probability generator whose probability distribution can be tuned as desired, and (7) a neuron, such as a McCullough-Pitts type neuron or perceptron.

We will implement most of the hardware with magnetics, as opposed to electronics, primarily because magnetic devices are usually (not always) more *energy-efficient* than their electronic counterparts and they are *non-volatile*. Both attributes are very desirable for edge computing and/or computing in resource-constrained environments (deep space, underground or underwater, etc.) where energy sources are scarce and all data must be stored in-situ in non-volatile elements because the cloud is either unavailable (deep space) or unreliable (enemy territory). The best designed hardware will comprise mostly magnetic elements with a smattering of electronic elements for gain, interfacing, high-speed and error-resilience, if needed.

We describe below the design of the seven essential elements.

A. The neuron

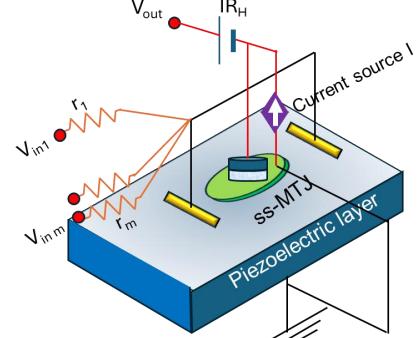
We will implement a McCullough-Pitts neuron or perceptron with a straintronic spin neuron (SSN) [10]. This model not only has its excellent energy efficiency and fast firing speed (sub-ns) but also keeps in line with our firing rate based synaptogenesis model. It relies on magnetization switching with electrically generated mechanical strain (hence the name “straintronic”).

The device is shown in Fig. 8(a) and consists of a ‘skewed straintronic magnetic tunnel junction’ (ss-MTJ) which has elliptical magnetic hard and soft layers whose major axes are *not* collinear (hence called ‘skewed’). The soft layer is in elastic contact with a piezoelectric thin film. The resistors $r_1 \cdot r_m$ encode synaptic weights. Because of any residual dipole coupling between the hard and the soft layer, the ss-MTJ is in the high resistance state in the absence of inputs. When the weighted sum of the input voltages $V_{in1}, V_{in2} \dots V_{inm}$ exceeds a certain threshold, the strain generated in the piezoelectric (and transferred to the soft layer) flips its magnetization and abruptly switches the resistance of the ss-MTJ from the high value R_H to the low value R_L , thereby switching the output voltage V_{out} from 0 to $I(R_L - R_H)$, where I is a bias current driven through the ss-MTJ. The working of this neuron was explained in [10] and hence omitted here. The output voltage V_{out} is given by [10]

$$V_{out} = f \left(\sum_i w_i V_{in_i} \right), \quad (13)$$

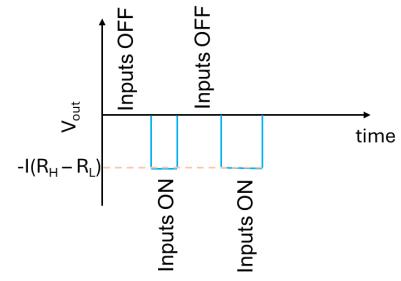
where f is our perceptron function and

$$w_i = \frac{r_1 \| r_2 \| \dots \| r_m}{r_1 \| r_2 \| \dots \| r_m + r_i}. \quad (14)$$



ss-MTJ = skewed straintronic magnetic tunnel junction

(a)



Firing sequence of the neuron

(b)

Figure 8: (a) Schematic of a straintronic spin neuron.
(b) The firing sequence of the neuron.

The firing sequence of the neuron is shown schematically in Fig. 8(b).

We can update the weights by updating the resistance r_i . Weights can be implemented with memristors whose resistance can be changed/updated with a (gate) voltage, but a better option is a simple straintronic magnetic tunnel junction (s-MTJ) [11] whose only difference with a generic magnetic tunnel junction is that the magnetization of the soft layer is rotated with electrically generated strain rather than the more common spin-transfer-torque, spin-orbit torque, or voltage-controlled-magnetic-anisotropy. The structure of a s-MTJ is shown in Fig. 9. The reason the s-MTJ is superior is that its transfer characteristic has a linear region where the conductance is *linearly proportional* to the gate voltage [12, 13] (see Fig. 10 for the computed transfer characteristic of a s-MTJ). Ref. [12] provided an analytical proof that such

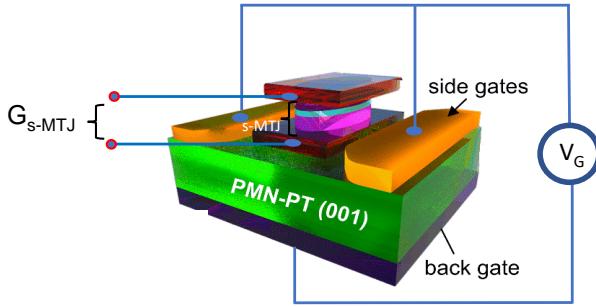


Figure 9: A straintronic magnetic tunnel junction (s-MTJ) reproduced with permission of the American Institute of Physics from [11]. The gate voltage V_G is applied across the piezoelectric PMN-PT between the (shorted) side gates and the back gate. This device exhibited a room-temperature tunneling magnetoresistance ratio (TMR) of slightly larger than 100%.

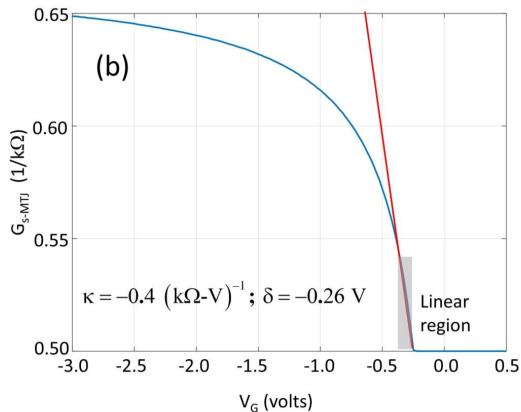


Figure 10: The transfer characteristic of a straintronic magnetic tunnel junction computed with the use of stochastic Landau-Lifshitz-Gilbert simulation of the soft layer's magnetodynamics in the presence of gate-voltage-generated strain at room temperature. This figure is reproduced from [12] with the permission of the Institute of Electrical and Electronics Engineers. The parameters for the soft layer were major axis = 800 nm, minor axis = 700 nm, thickness = 2.2 nm, saturation magnetization $M_s = 8.5 \times 10^5 \text{ A/m}$, dipole coupling field $H_d = 1000 \text{ Oe}$, Gilbert damping constant = 0.1, saturation magnetostriction $\lambda_s = 600 \text{ ppm}$, Young's modulus $Y = 120 \text{ GPa}$, piezoelectric coefficient $d_{33} = 1.5 \times 10^{-9} \text{ C/N}$ and the piezoelectric layer thickness $t = 1 \mu\text{m}$. The value of R_{AP} was assumed to be 2 k Ω and the value of R_P was 1 k Ω .

a linear region must exist and that its width is proportional to the antiparallel resistance of the s-MTJ. This

linearity is very desirable and not easily available in any other device such as a memristor. The linearity allows more accurate control of the conductance where equal increments or decrements of the gate voltage will result in equal increase (long term potentiation) or decrease (long term depression) of the conductance improving accuracy in classification tasks [14, 15]. It may also benefit adaptive synaptogenesis.

B. Non-volatile weights

Unlike memristor based weights, weights implemented with s-MTJs may appear to be *volatile* since the gate voltage should be kept on to maintain the conductance/resistance of the s-MTJ at a particular value. However, it must be noted that all that the gate voltage does is produce a strain in the s-MTJ's soft layer. If the produced strain survives after the gate voltage is removed, then the gate voltage need not be kept on and the weight would be non-volatile. This obviously requires *non-volatile strain*. Non-volatile strain induced by an electrical voltage in a piezoelectric has been reported by a number of authors [16–19]. At this time, however, the robustness of non-volatile strain is questionable, i.e., whether it survives thermal recycling, and the retention time is undetermined, but the fact that remanent strain is found after withdrawal of the voltage that induces strain in a piezoelectric is a very encouraging development which portends *non-volatile weights* when they are implemented with s-MTJs.

C. A device to measure the firing rate of a neuron

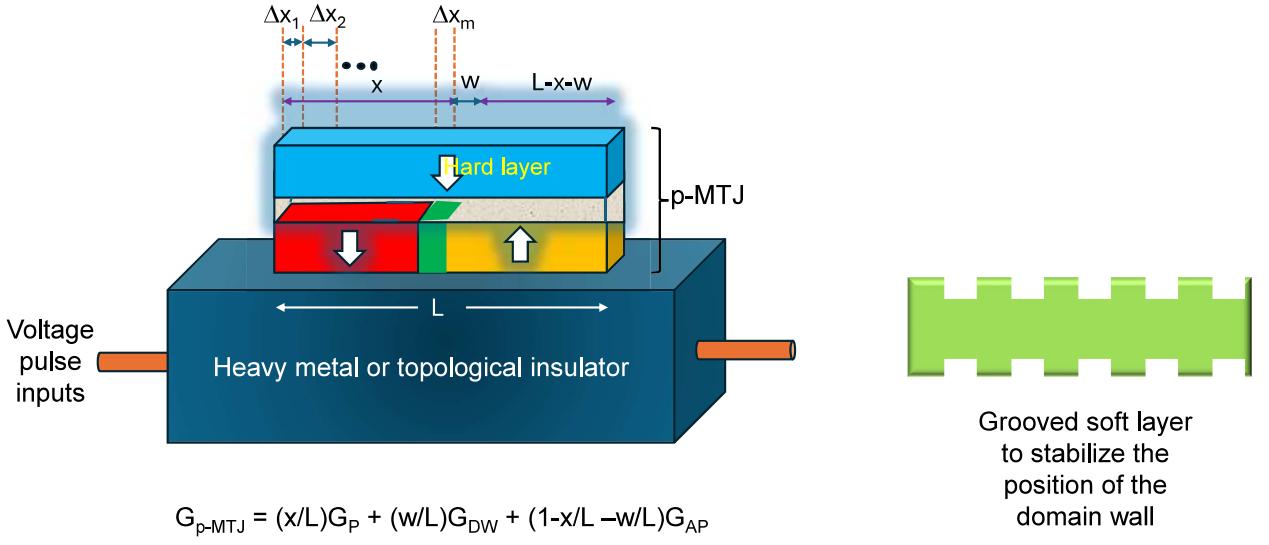
To measure the firing rate of the neuron, we employ a domain wall synapse (DWS) [20, 21] shown in Fig. 11. It consists of a p-MTJ where the ferromagnetic hard and soft layers have perpendicular magnetic anisotropy. The p-MTJ is fabricated on a heavy metal layer or a topological insulator. The output voltage pulses generated by a firing neuron are converted to current pulses and injected into the heavy metal layer or topological insulator to create successive pulses of spin-orbit torque that move the domain wall in the soft layer of the p-MTJ by discrete amounts Δx_n .

The conductance of the p-MTJ is given by [20]

$$G_{p-MTJ} = \frac{x}{L} G_P + \frac{W}{L} G_{DW} + \left(1 - \frac{x+W}{L}\right) G_{AP}, \quad (15)$$

where G_P is the parallel conductance and G_{AP} is the antiparallel conductance of the p-MTJ, G_{DW} is the conductance associated with the domain wall, L is the length of the soft layer, and $x = \sum_i^m \Delta x_i$, the total domain wall displacement after m pulses.

In order to measure the firing rate, we build the circuit of Fig. 12 where the DWS is placed in series with a



DOMAIN WALL SYNAPSE (DWS)

Figure 11: A domain wall synapse. The domain wall position can be stabilized in the presence of thermal noise by using a periodically grooved soft layer

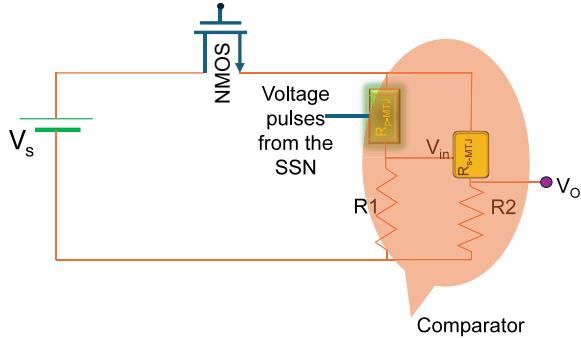


Figure 12: Circuit to measure the firing rate and compare it with a threshold.

resistor R_1 and connected to a power supply voltage V_s through an NMOS transistor. The gate voltage of the NMOS is turned on for a time duration T and let us assume that the SSN fires m times during that period. Because of slight dipole coupling between the hard and the soft layers of the p-MTJ, it will initially be in the antiparallel state. After time T , the conductance of the

p-MTJ will be

$$\begin{aligned} G_{p-MTJ} &= \frac{\sum_i^m \Delta x_i}{L} G_P + \frac{W}{L} G_{DW} \\ &\quad + \left(1 - \frac{\sum_i^m \Delta x_i + W}{L}\right) G_{AP} \\ &= m \frac{\Delta x}{L} G_P + \frac{W}{L} G_{DW} + \left(1 - \frac{\Delta x + W}{L}\right) G_{AP} \\ &= m \frac{\Delta x}{L} (G_P - G_{AP}) + \frac{W}{L} G_{DW} \\ &\quad + \left(1 - \frac{W}{L}\right) G_{AP}, \end{aligned} \quad (16)$$

where Δx is the average displacement of the domain wall after a pulse. It can be determined experimentally with a magnetic force microscope after the soft layer has been subjected to a number of pulses. Therefore,

$$\begin{aligned} m &= \frac{L}{\Delta x} \frac{1}{(G_P - G_{AP})} \times \\ &\quad \left[G_{p-MTJ} - \frac{W}{L} G_{DW} - G_{AP} \left(1 - \frac{W}{L}\right) \right]. \end{aligned} \quad (17)$$

The quantities W , L , G_P , G_{AP} , G_{DW} and Δx are known. Hence by measuring G_{p-MTJ} , one can deduce the value of m . The firing rate is simply m/T . Thus, one can measure the firing rate.

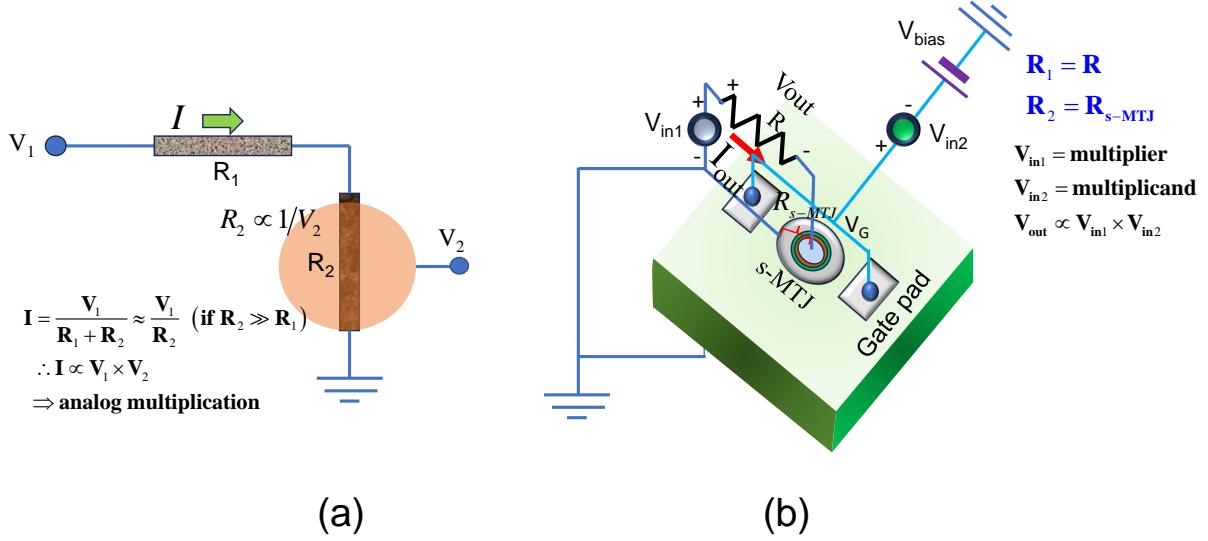


Figure 13: (a) A voltage divider with a voltage tunable resistor can implement an **analog multiplier**. (b) An actual implementation of an analog multiplier with a s-MTJ biased in the linear region of the transfer characteristic and some bias sources. Reproduced from [22] with permission of the Institute of Electrical and Electronics Engineers.

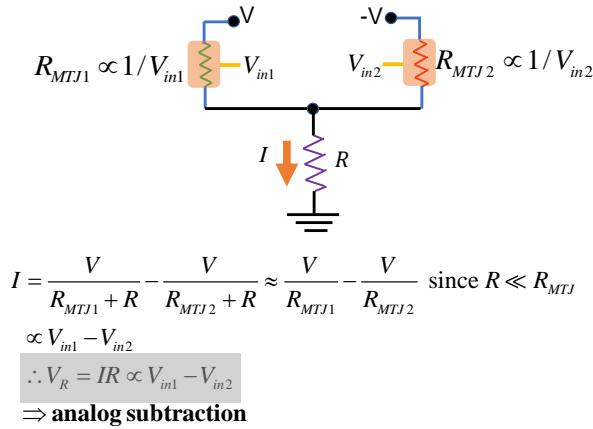


Figure 14: An **analog subtractor** implemented with two s-MTJs acting as resistors whose resistances are inversely proportional to the gate voltages applied to them.

D. Determining if the firing rate exceeds a threshold - the comparator

From Equation (16), we see that the p-MTJ conductance is linearly proportional to m and hence to the firing rate m/T . Let the p-MTJ conductance corresponding to a threshold firing rate be G_{p-MTJ}^T and its reciprocal be the resistance R_{p-MTJ}^T . In the voltage divider circuit of Fig. 12, the voltage V_{in} is $\frac{R_1}{R_1 + R_{p-MTJ}}$. At the threshold

value,

$$V_{in}^T = \frac{R_1}{R_1 + R_{p-MTJ}^T} V_s. \quad (18)$$

In Fig. 12, the s-MTJ is a large cross-section strain-tronic MTJ whose resistance switches somewhat abruptly from high to low when the voltage applied to its gate exceeds a threshold value [11]. The threshold depends on the energy barrier within the soft layer of the s-MTJ and hence can be engineered by engineering the shape of the s-MTJ (the ellipticity of the soft layer). We design the s-MTJ such that the threshold voltage for switching is equal to V_{in}^T . Then, if the firing rate of the SSN exceeds the threshold, the s-MTJ resistance will switch from high to low. Since $V_O = R_2 / (R_2 + R_{s-MTJ})$, a high value of V_O will imply that the threshold firing rate has been exceeded, and a low value will indicate that the threshold has not been reached. Thus, by monitoring V_O , we can infer if the firing rate has exceeded a threshold. This is the basis of the comparator.

E. Analog multiplier

An analog multiplier can be realized with a simple voltage divider network consisting of two resistors, one of whose resistance can be tuned with a gate voltage such that resistance is inversely proportional to the gate voltage. The implementation is shown in Fig. 13. The voltage-dependent resistor R_2 is obviously implemented with a s-MTJ biased in the linear region of the transfer characteristic shown in Fig. 10 where the conductance is proportional to the applied gate voltage and hence the

resistance is inversely proportional to the gate voltage. An exact implementation of the analog multiplier can be found in Ref. [12, 22] and hence not repeated here.

F. Analog subtractor

An analog subtractor can also be realized with two voltage divider networks where once again we use s-MTJs to implement resistors whose resistances are inversely proportional to the gate voltage. We show only the circuit representations in Fig. 14 since the device implementation follows trivially from there.

G. A probability generator randomly generating binary bits 0 and 1 with tunable distribution of the probabilities of either bit

This device can be realized with a simple binary stochastic neuron (BSN) implemented with a low barrier nanomagnet. The low barrier nanomagnet forms the soft layer of an MTJ whose resistance encodes the magnetization orientation of the soft layer. The resistance is given by

$$R_{MTJ} = R_P + \frac{R_{AP} - R_P}{2} [1 - \cos\theta], \quad (19)$$

where $R_{P(AP)}$ is the parallel (antiparallel) resistance of the MTJ and θ is the angle between the magnetizations of the hard and the soft layer. If the magnetization of the soft layer subtends an acute angle with that of the hard layer, then the resistance is low and is interpreted as bit 0, whereas if the angle is obtuse, the resistance is high and is interpreted as bit 1. The probability of obtaining the bit zero can be tuned by injecting a spin polarized current into soft layer with varying spin polarization.

Let us say that the unit vector along the magnetization of the hard layer is \hat{m} and that along the spin polarization is \hat{s} . If the scalar product $\hat{m} \cdot \hat{s}$ is positive, we will call the spin polarized current positive; otherwise, we will consider it as negative. The probability of the measured bit being zero depends on the sign and magnitude of the spin polarized current

V. CONCLUSION

Adaptive synaptogenesis has been discussed as a neuromorphic learning approach that mitigates the challenge of catastrophic forgetfulness inherent to conventional ANN retraining. Rather than overwriting all past memory in a single disruptive event, adaptive synaptogenesis enables constructive learning over time and efficient

long-term memory retention. Adaptively and continually adding and culling synaptic connections in an artificial neural network retains long-term learning gleaned through longitudinal experience. Conventional, global retraining approaches are inherently incapable of such comprehensive subjective experiential learning that encompasses even temporally distant or sparse events. The adaptive synaptogenesis approach is particularly relevant for edge sensing applications, in which data connectivity, bandwidth limitations, or computing resource constraints make complete periodic retraining impractical or impossible.

Simulation models demonstrated the functional viability of this construct, both in supervised and unsupervised modes of learning. Model inputs correlated with classes were automatically encoded through the creation of new synaptic connections and unused or spurious connections were shed for computational efficiency. Learning accuracy for adaptive synaptogenesis modes were comparable to kNN in a series of benchmark tests.

Beyond the theoretical construct and modeling of this architecture, a hardware implementation has been proposed using power-efficient nanomagnetic devices, capable of variable output that can be changed within a nanosecond. These straintronic spin neurons use linearly gate-regulated channel conductance to encode weights in memory. While nonvolatility of strain state has been reported, more work is required to understand the environmental limits. Furthermore, several logic devices have been described as functional elements of an ANN with adaptive synaptogenesis capability, including a firing rate comparator, multiplier, and subtractor.

In future work, the adaptive synaptogenesis model will be implemented using straintronic magnetic tunnel junction devices and characterized for comparison to the simulation results reported herein. Subsequently, the hardware will be used for real-time learning applications, such as embedded classification techniques in sensing applications and fluid neuromuscular control in robotic applications.

ACKNOWLEDGMENTS

The authors are indebted to our late colleague Prof. William Levy of the University of Virginia for many illuminating discussions pertaining to neuroscience. This project was funded in part by the US National Science Foundation IUCRC Multifunctional Integrated System Technology Center at University of Virginia, US Air Force Office of Scientific Research under grant FA865123CA023 given to the iDISPLA Convergence Lab at Virginia Commonwealth University and by the Virginia Innovation Partnership Corporation grant CCF23-0114-HE.

-
- [1] H. J. S. Feder, J. J. Leonard, and C. M. Smith, Adaptive mobile robot navigation and mapping, *The International Journal of Robotics Research* **18**, 650 (1999).
- [2] Y. Mei, Y.-H. Lu, Y. Hu, and C. Lee, Deployment of mobile robots with energy and timing constraints, *IEEE Transactions on Robotics* **22**, 507 (2006).
- [3] I. Lluvia, E. Lazcano, and A. Ansueategi, Active mapping and robot exploration: A survey, *Sensors* **21**, 10.3390/s21072445 (2021).
- [4] M. Mermilliod, A. Bugaiska, and P. BONIN, The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects, *Frontiers in Psychology* **4**, 10.3389/fpsyg.2013.00504 (2013).
- [5] E. Hajizada, P. Berggold, M. Iacono, A. Glover, and Y. Sandamirskaya, Interactive continual learning for robots: a neuromorphic approach, in *International Conference on Neuromorphic Systems 2022*, ICONS '22 (ACM, New York, NY, USA, 2022).
- [6] W. B. Levy, A sequence predicting ca3 is a flexible associator that learns and uses context to solve hippocampal-like tasks, *Hippocampus* **6**, 579 (1996).
- [7] D. A. August and W. B. Levy, Temporal sequence compression by an integrate-and-fire model of hippocampal area ca3, *Journal of Computational Neuroscience* **6**, 71 (1999).
- [8] D. Hendrycks and K. Gimpel, Gaussian error linear units (gelus), arXiv preprint arXiv:1606.08415 (2016).
- [9] R. A. Baxter and W. B. Levy, Constructing multilayered neural networks with sparse, data-driven connectivity using biologically-inspired, complementary, homeostatic mechanisms, *Neural Networks* **122**, 68 (2020).
- [10] A. K. Biswas, J. Atulasimha, and S. Bandyopadhyay, The straintronic spin neuron, *Nanotechnology* **26**, 285201 (2015).
- [11] Z. Y. Zhao, M. Jamali, N. D'Souza, D. Zhang, S. Bandyopadhyay, J. Atulasimha, and J. P. Wang, Giant voltage manipulation of mgo-based magnetic tunnel junctions via localized anisotropic strain: A potential pathway to ultra-energy-efficient memory technology, *Appl. Phys. Lett.* **109**, 092403 (2016).
- [12] R. Rahman and S. Bandyopadhyay, A non-volatile all-spin non-binary matrix multiplier: An efficient hardware accelerator for machine learning, *IEEE Trans. Elec. Dev.* **69**, 7120 (2022).
- [13] S. Bandyopadhyay, Straintronic magnetic tunnel junctions for analog computation: A perspective, *J. Phys. D: Appl. Phys.* **58**, 152001 (2025).
- [14] R. S. Yadav, P. Gupta, A. Holla, K. I. A. Khan, P. K. Muduli, and D. Bhowmik, Demonstration of synaptic behavior in a heavy metal ferromagnetic metal oxide heterostructure based spintronic device for on-chip learning in crossbar-array-based neural networks, *ACS Appl. Electron. Mater.* **5**, 484–97 (2023).
- [15] V. B. Desai, D. Kaushik, J. Sharda, and D. Bhowmik, On-chip learning of a domain-wall-synapse-crossbar-array-based convolutional neural network, *Neuromorph. Comput. Eng.* **2**, 024006 (2022).
- [16] A. Chen and et al., Giant nonvolatile manipulation of magnetoresistance in magnetic tunnel junctions by electric fields via magnetoelectric coupling, *Nat. Commun.* **10**, 243 (2019).
- [17] L. Yang, Y. Zhao, S. Zhang, P. Li, Y. Gao, Y. Yang, H. Hu, P. Miao, Y. Liu, A. Chen, C. W. Nan, and C. Gao, Bipolar loop-like non-volatile strain in the (001)-oriented pb_{(mg_{1/3}nb_{2/3})₀₃}-pb₂o₃ single crystals, *Sci. Rep.* **4**, 4591 (2014).
- [18] T. Wu and et al., Domain engineered switchable strain states in ferroelectric (011) [pb(mg_{1/3}nb_{2/3})₀₃]_{1-x}-[pb₂o₃]_x (pmn-pt, x=0.32) single crystals, *J. Appl Phys.* **109**, 124101 (2011).
- [19] T. Wu and et al., Electrical control of reversible and permanent magnetization reorientation for magnetoelectric memory devices, *Appl. Phys. Lett.* **98**, 262504 (2011).
- [20] A. Sengupta, Y. Shim, and K. Roy, Proposal for all-spin artificial neural network: Emulating neural and synaptic functionalities through domain wall motion in ferromagnets, *IEEE Trans. Biomed. Circ. Syst.* **10**, 1152–1160 (2016).
- [21] S. Liu, I. P. Xiao, C. Cui, J. A. C. Incorvia, C. H. Bennett, and M. J. Marinella, A domain wall magnetic tunnel junction artificial synapse with notched geometry for accurate and efficient training of deep neural networks, *Appl. Phys. Lett.* **118**, 202405 (2021).
- [22] S. Bandyopadhyay, Magnetic straintronics for ultra-energy-efficient unconventional computing: A review, *IEEE Trans. Magn.* **60**, 4100110 (2024).