

# A Low-Power Bayesian Head Using SOT-MRAM Arrays for Uncertainty-Aware Binary Neural Networks

Joao Henrique Quintino Palhares  
SPINTEC, CEA  
Grenoble, France

Joao-  
henrique.QUINTINOPALHARES@cea.fr

Bruno Lovison Franco  
Université de Montpellier, LIRMM  
Montpellier, France

bruno.lovison-franco@lirimm.fr

Louis Hutin  
LETI, CEA  
Grenoble, France

jonathan.miquel@grenoble-inp.fr

Jonathan Miquel  
INP grenoble, SPINTEC, CEA  
Grenoble, France  
jonathan.miquel@grenoble-inp.fr

Kamel-eddine Harabi  
LETI, CEA  
Grenoble, France  
kamel-eddine.harabi@cea.fr

Aymen Romdhane  
SPINTEC, CEA  
Grenoble, France  
aymen.romdhane@cea.fr

Kevin Garello  
SPINTEC, CEA  
Grenoble, France  
kevin.garello@cea.fr

**Abstract**—This work presents a novel low-power, mixed-signal computing in memory (CIM) architecture for Bayesian inspired inference, targeting edge AI applications requiring energy-efficient uncertainty estimation. Our system integrates a deterministic Binary Neural Network (BNN) with a Bayesian head module implemented using multi-pillar (MP) Spin-Orbit Torque Magnetic RAM (SOT-MRAM) based arrays. The Bayesian head perturbs the output popcount of the BNN by injecting configurable stochastic counts, enabling uncertainty quantification in classification tasks. These perturbations are configurable in ‘flavor’ through a tunable dropout rate and the number of MP cells. A VCO-based ADC converts analog resistive summations into digital counts, which are then combined with the deterministic BNN output. On MNIST and CIFAR-10, the proposed system achieves classification accuracy comparable to state-of-the-art Bayesian approaches while consuming only 19  $\mu$ W. It achieves a favorable energy efficiency of 53 TOPs/W (18.9 fJ/OPS) for 3-bits and 110 TOPs/W for 2-bits perturbation precision. Uncertainty estimation is validated through controlled domain shifts (e.g., tilted images), showing robust entropy and variance evolution. Notably, the proposed uncertainty estimation requires only 25 perturbation runs, resulting in a total energy cost of just 454 fJ. At this overhead, the Bayesian-inspired model improves reliability by 34.29% compared to the baseline on CIFAR-10. This low-power hybrid analog-digital architecture offers a promising solution for edge applications with embedded confidence metrics.

**Keywords**—Bayesian approximation, Multi-pillar SOT MRAMs, Bayesian Binary Neural Networks, Uncertainty estimation, computing in memory.

## I. INTRODUCTION

Edge intelligence demands neural network architectures that are not only energy-efficient but also capable of quantifying predictive uncertainty. Bayesian Neural Networks (BayesNNs) offer an interesting approach to estimate uncertainty by capturing prediction confidence through probabilistic inference[1]. However, conventional Bayesian methods are typically resource-intensive and not suitable for energy-constrained environments[2].

In this work, we propose a mixed-signal Bayesian head (BH) computing-in-memory (CIM) architecture implemented in SOT-MRAM arrays. The BH approximates Bayesian

inference in a deterministic Binary Neural Network (BNN) by introducing tunable stochastic perturbations to the partial sums (reducing implementation overhead). Unlike prior bayesian approximations such as Monte Carlo Dropout[3], which injects stochasticity via randomly masking activations, or Bayes by Backprop[4], [5], which learns distributions over the weights[6], our approach introduces empirically calibrated perturbations only at inference time, avoiding retraining. Importantly, in our evaluation, these perturbations are injected post-training during inference, following a Bayesian inspired predictive averaging strategy rather than embedding stochasticity directly in training. This enables uncertainty estimation without retraining and reduces implementation overhead[7], [8]. The system combines multilevel, multi-pillar (MP) SOT-MRAM memory [9], with time-to-digital converter (TDC), voltage-controlled oscillator (VCO) based ADCs to realize in-situ uncertainty-aware inference. We evaluate the architecture through circuit-level simulations and algorithmic modeling using the BinaryNET model in a VGG based architecture on CIFAR-10 dataset and LarqNET on MNIST [10], [11]. Uncertainty estimation is assessed under domain shift conditions (rotated images), and energy efficiency (TOPs/W) is compared against existing bayesian hardware implementations.

## II. BAYESIAN METHOD APROXIMATION FRAMING

In Bayesian deep learning, the predictive distribution for a class  $y$  given an input  $x$  and dataset  $D$  is:

$$p(y|x,D) = \int p(y|x,w)p(w|D)dw \quad (1)$$

Where  $p(w|D)$  represents the posterior over network weights given training data  $D$ , and  $p(y|x,w)$  is the likelihood of the prediction given weights  $w$ . Once computing this integral is intractable for deep neural networks, approximate inference methods such as Monte Carlo Dropout, variational inference, are typically used to replace the true posterior with a tractable surrogate [3], [4], [5]. Prior works have explored learned posterior approximations and multiplicative noise injection (ScaleDropout). [12], [13] to model predictive uncertainty. In contrast, we introduce an additive perturbation mechanism applied to deterministic BNN outputs, specifically designed for seamless integration with binary CIM architectures, where

inference is performed via popcount accumulation. This enables perturbations to be directly injected into the output counts of the array while preserving the core computation pipeline.

Our method uses a hardware-friendly Bayesian head (BH) module that injects stochastic perturbations  $P$  into a deterministic output  $Y$  of a Binary Neural Network (BNN). The BNN serves as a point-estimate baseline with trained weights  $w_{\text{BNN}}$ , approximating a maximum a posteriori (MAP) solution. To emulate posterior variability, perturbations are applied to the intermediate BNN outputs during inference:

$$\widehat{Y}_k = Y_k + m_k P_k \quad (2)$$

Where  $Y_k$ , is the deterministic output of column  $k$ ,  $m_k$  Bernoulli ( $p$ ) is a dropout like mask,  $p_k$  is a perturbation sample from an empirically calibrated distribution  $P(\theta)$ , and  $\widehat{Y}_k$  the perturbed version of the deterministic output.  $P(\theta)$  is configured to approximate the uncertainty behavior observed in bayesian inference when the input distribution shifts (out of domain data). In our crossbar-based implementation, these  $Y_k$  correspond to column-wise partial sums, as detailed in section III and illustrated in Fig. 1.  $P(\theta)$  calibration is discussed in section IV and the uncertainty behavior in section V. The overall predictive distribution is approximated via Monte Carlo sampling:

$$p(y|x, D) \approx \frac{1}{T} \sum_{t=1}^T f(x; w_{\text{BNN}}, m^t, P^t) \quad (3)$$

Each sample uses a different mask  $m$  and perturbation  $P$  yielding a tractable surrogate for the Bayesian predictive distribution without retraining. This leads to a variational approximation to the Bayesian predictive distribution:

$$\tilde{p}(y|x) = \int p(y|x, w_{\text{BNN}} + P) q(P) dP \quad (4)$$

Where  $\tilde{p}$  is the approximate predictive distribution and  $q(P)$  the surrogate distribution approximating the effect of sampling from the true posterior distribution over weights or predictions. In our evaluation, the surrogate is not optimized during training but calibrated empirically at inference time, in order to implement a Bayesian inspired post hoc uncertainty estimation [7], [8] rather than fully Bayesian training.

### III. SYSTEM ARCHITECTURE OVERVIEW

The architecture consists of two components: a deterministic BNN inference core and a stochastic BH, both implemented using SOT-MRAM-based crossbar arrays (Fig. 1). The deterministic BNN performs binary XNOR operation and popcount operations to compute the partial sum, while the BH module introduces controlled perturbation post-accumulation to emulate bayesian sampling.

The BNN crossbar array handles up to 64 binary inputs, yielding a maximum deterministic partial sum ( $y_i$ ) of 64. To reflect uncertainty, the BH injects bounded perturbations ( $P_i(E)$ ) (up to 7 counts, 3-bit resolution,  $\sim 10\%$  of the deterministic sum). These are applied stochastically across multiple runs to generate a prediction distribution from which confidence is estimated (variance). The BH itself consists of up to 16 serially connected multi-pillar (MP) SOT-MRAM devices per array column, enabling a wide equivalent resistance range. The BH itself consists of up to 16 serially connected multi-pillar (MP) SOT-MRAM devices per array

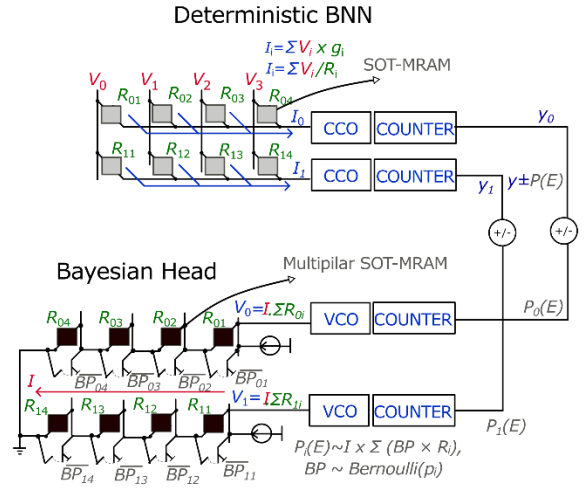


Fig. 1: Simplified schematics of the deterministic BNN and BH MRAM arrays. The BNN uses XNOR-based MAC operations with current-to-count conversion via a CCO ADC. The BH injects signed perturbation counts from randomly configured multi-pillar (MP) SOT-MRAM chains, digitized by a VCO ADC. The Bayesian BNN output combines deterministic counts  $Y$  with random counts  $P(E)$  over multiple runs.

column, enabling a wide equivalent resistance range. This structure provides fine grained control over the total equivalent resistance of the MP array column, ensuring sufficient resolution for the quantization by the VCO based ADC stage. Each MP is conditionally bypassed by a parallel-

connected transistor, whose gate is driven by a control bit (BP) generated by an 8-bit Linear Feedback Shift Register (LFSR). This LFSR produces dropout-like stochastic masks based on programmable dropout rates ranging from 20% to 90%.  $R_{\text{EQ}}$  is read at a constant current, and the resulting reading voltage is used for further digitalization. The role of SOT-MRAM in the proposed architecture is not to rely on intrinsic device noise, but to provide a reconfigurable ensemble of resistive summation whose composition can be probabilistically modulated by the control bit BP. The uncertainty thus emerges from controlled hardware variability rather than uncontrolled physical noise. This provides the resulting perturbation distribution  $P(E)$  that is randomly added or subtracted to the deterministic partial sum  $y_i$ :

$$P(E)_i \sim I(\sum \overline{BP}_i R_{ij}) \quad (5)$$

$P(E)$  is the perturbation distribution output of the BH,  $I$  is the current passing through the MP array,  $BP$  the by-pass switch control signal, and  $R_{ij}$  the resistance of the  $i$ - $j$  MP.

### IV. CIRCUIT IMPLEMENTATION AND SIMULATIONS

Simulations were carried out using Cadence Spectre and GlobalFoundries 22 nm FDX PDK. The BH control logic, including the LFSR-based dropout generator, was synthesized using Synopsys Design Compiler and standard cell libraries from the same technology node. The SOT-MRAM array was modeled using a verilogA behavioral model[14].

#### A. MP SOT-MRAM

Spin-Orbit Torque Magnetic RAM (SOT-MRAM) is a non-volatile memory in which data is stored in the resistance state of a Magnetic Tunnel Junction (MTJ). The MTJ comprises two ferromagnetic layers, a fixed (pinned) layer (PL) and a free layer (FL), separated by a thin oxide barrier (see Fig. 2a). Its resistance state is based on the relative



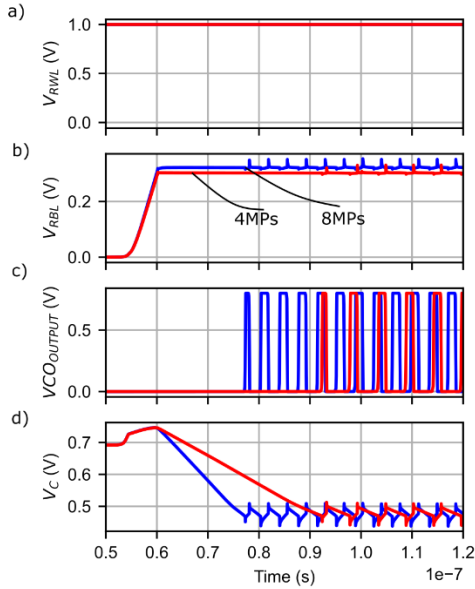


Fig. 4: transient analysis of the VCO. a)  $V_{RWL}$  activation, b) voltage drop across BL -BSL,  $V_{RBL}$ . c) The corresponding VCO output and the VC charge and discharge dynamics according to  $R_{EQ}$  and the number of MPs.

In our design, a 3-bit counter records up to 7 pulses per sampling window. This count resolution enables perturbation levels that span at least 10% of the deterministic BNN’s partial sum, which can reach a maximum of 64 counts per column in a  $64 \times 64$  crossbar. Although only 3 bits (8 levels, up to 7 counts) are used to digitize the perturbation, the BH array consists of 16 MPs (rows), resulting in a more continuous and well resolved analog frequency spectrum before quantization improving the quality of the sampled perturbation distribution. The integration time  $t$ , or sampling window, is defined by the ratio of the counter’s maximum resolution to the VCO’s peak frequency:

$$t = \frac{(2^n - 1)}{f_{MAX}} \quad (6)$$

$$\text{Counts} = \lfloor f_i \cdot t \rfloor \quad (7)$$

$t$  is the integration time per sampling,  $n=3$  for the 3-bit counter,  $f_{MAX}$  is the maximum output frequency of the VCO and  $f_i$  is the instantaneous frequency corresponding to the sampled resistance state. The maximum frequency is roughly 439 Mhz yielding a sampling time of 15.9 ns. The count score spanning the whole range of the BH array (from 0 to 16 MPs) is shown in Fig 5b.

### E. Noise Flavor Control

A key strength of the BH is the ability to tune its output perturbation distribution, referred to as the ‘noise flavor’. It can be tuned by adjusting the dropout rate of the BP transistors and the number of MPs configured within the BH array. This tunability provides control over not only the entropy and bias of the injected perturbation but also its mean ( $\mu$ ), standard deviation ( $\sigma$ ), and overall distribution shape. Fig. 6a shows a lookup table (LUT) mapping different dropout configurations to output count distributions, where the inset values represent the mean ( $\mu$ ) and the colorbar indicates the standard deviation ( $\sigma$ ). By increasing the number of MPs, both  $\mu$  and  $\sigma$

rise due to the additive contribution of multiple MP cells. Conversely, increasing the dropout rate reduces  $\mu$  and  $\sigma$  by deactivating active perturbation paths. This two-dimensional

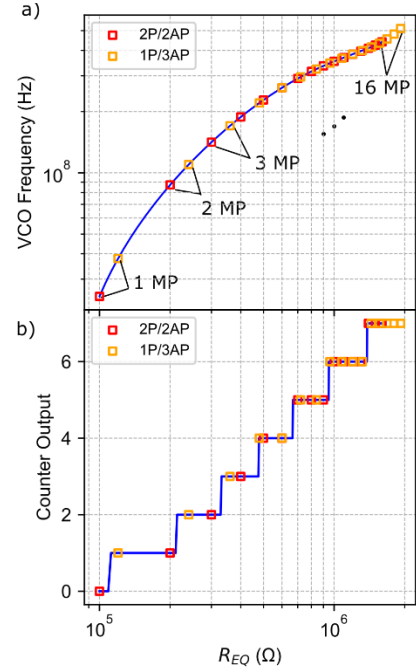


Fig. 5: a) VCO transfer function showing output frequency as a function of total equivalent resistance ( $R_{EQ}$ ) for up to 16 MPs across BL–SL (i. e., set to 2AP/2P (in red) and 1P/3AP (in orange)). The frequency ranges from 24.7 MHz to 440 MHz. (b) Corresponding 3-bit counter output (count score) versus  $R_{EQ}$ .

parameter space enables flexible control over the injected uncertainty magnitude and shape which is bound to a fraction of the BNN partial sum. In practice, although the LUT is technology-dependent, it is generated through circuit-level characterization, and these parameters are calibrated offline once per technology corner and remain fixed during inference (similarly to ADC calibration in mixed-signal CIM systems).

Fig. 6b illustrates how these configurations impact the shape of the perturbation distribution for 16 MPs under dropout rates ranging from 20 % to 90 %, 20 % (i), 50 % (ii), and 90 % (iii). At the extremes conditions, towards 20% and 90% dropout rates and maximum number of MPs (i), the distribution resembles a half-normal centered at (i) zero or (iii) clipped at maximum number of counts. At moderate configurations (ii), the distribution resembles a more symmetric, normal-like gaussian shape (highlighted in yellow in Figure 6a). These profiles enable emulation of varying uncertainty behaviors seen in bayesian inference, aligning with flexible confidence calibration. In our simulation framework, the bayesian inspired method for the BNN employs a half-normal-like distribution centered at zero with random signed addition and subtraction of perturbation, as illustrated in Fig. 6b-iii. This shape is selected aligning with the chosen operation of random addition and subtraction of the perturbation in our hardware-aware simulation. Other configurations, such as normal-like or max-centered half-normal distributions, and alternative operations (e.g., additive biasing, subtractive shifts, or multiplicative scaling), remain unexplored and may extend the applicability of the BH to approximate more complex posterior[3], [4], [5], [13].

## V. SIMULATION OF UNCERTAINTY ESTIMATION AND ROBUSTNESS

To assess the effect of our Bayesian Head (BH) perturbation on predictive reliability, we conduct inference-time simulations using the BinaryNET model implemented in

a VGG-based architecture for the CIFAR-10 dataset and LarqNET for MNIST. BinaryNET is a binarized convolutional neural network with all weights and activations constrained to  $\pm 1$ , except for the first and last layers, which remain in full precision to improve overall accuracy [10], [11].

Following the variational surrogate framing introduced in Section II, the deterministic BNN acts as a point-estimate predictor (baseline), while the BH module emulates posterior sampling by injecting stochastic perturbations into the intermediate layer partial sums. The  $64 \times 64$  crossbar array is used to map partial sums across the hidden layers, excluding the first and last layers. The mapping follows the scheme in [16] where each column of the array represents a  $K \times K \times D$  kernel.  $K$  is the kernel size and  $D$  the input feature depth. Each partial sum output is perturbed as shown in equation (2). Here,  $Y_k$  is the deterministic popcount output for kernel  $k$ ,  $P_k$  is a perturbation sampled from a calibrated distribution controlled by BH parameters (dropout rate and MP count), and  $mk$  is a binary mask that decides whether the perturbation is applied. Importantly, Note that  $mk$  is distinct from the mask BP used to configure the BH array itself: it operates after sampling  $P_k$  to decide whether the perturbation is injected (considered in the simulation).

This stochastic process effectively samples from a variational surrogate distribution, approximating Bayesian predictive inference without retraining the BNN. Perturbations are applied only at inference time and are selected empirically from a pre-characterized set of distributions in the BH LUT. To account for hardware effects, the deterministic partial sums already incorporate process variability (e.g., memory array access transistors and ADCs). This inherent noise makes more difficult the separation between in-domain and out-of-domain data, but at the same time strengthens the robustness of our Bayesian approximation.

In practice, multiple candidate perturbation distributions from the LUT are tested, and the one yielding the most consistent joint behavior of predictive accuracy and variance under shifted data (rotated images) is retained for use. Our approach requires only 25 perturbation injections per inference to estimate uncertainty, far less than typical Bayesian or ensemble-based methods, making it both low-power and efficient (see Section VI). This ‘‘Bayes-inspired’’ strategy contrasts with fully Bayesian or variational methods, where uncertainty is embedded during training. Instead, we inject the calibrated perturbation distribution post-training (post hoc dropout-injection method), like a stochastic readout head, providing predictive averaging without modifying the deterministic training pipeline [7], [8]. To validate the uncertainty estimation behavior, we apply controlled domain shift by rotating CIFAR-10 and MNIST test images at increasing angles. The image rotation is used to introduce input mismatch while preserving class labels, enabling controlled evaluation of uncertainty.

Figure 7 summarizes the impact of domain shift on accuracy, uncertainty, and filtering. In this analysis, we apply an uncertainty filter that discards predictions whose variance exceeds a threshold derived from the empirical training distribution (third quartile). The goal is to retain only confident predictions, thereby improving reliability at the cost of coverage. As shown in Fig. 7a–b, predictive accuracy decreases progressively as the rotation angle increases, reflecting the growing mismatch between shifted inputs and

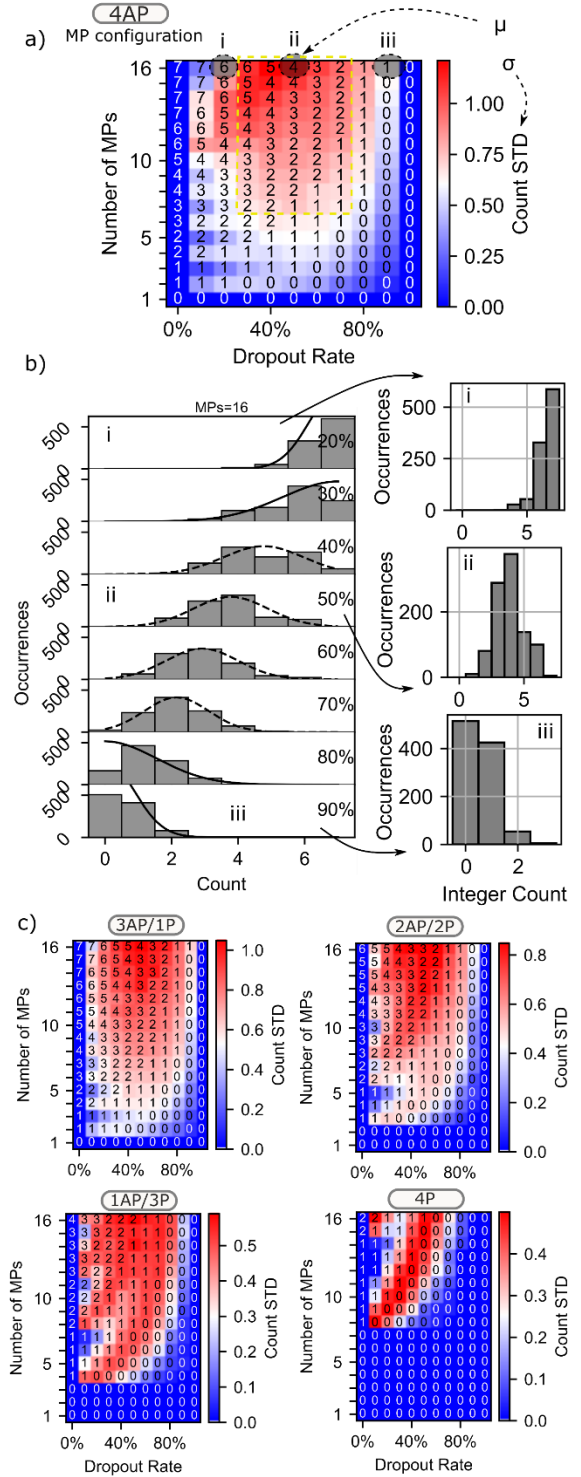


Fig. 6: a) The Bayesian Head (BH) lookup table (LUT), or count perturbation space, controlled by the dropout rate and the number of MPs at 2AP/2P configuration. b) Count distributions for 16 MPs at dropout rates ranging from 20% to 90%. 20% (i), 50% (ii), and 90% (iii). c) LUT for 3AP/1P, 2AP/2P, 1AP/3P and 4P configurations.

the training data. Accuracy drops from 0.90 to 0.35 on CIFAR-10 and from 0.96 to 0.18 on MNIST. When considering only the retained predictions, however, the accepted accuracy improves substantially: on average, it is 34.29% higher for CIFAR-10 and 9.61% higher for MNIST compared to the raw accuracy, with the maximum accuracy for CIFAR-10 increasing from 0.90 to 0.98. Fig. 7c–d shows that predictive variance rises consistently as rotation grows. This indicates reduced model confidence under domain shift,

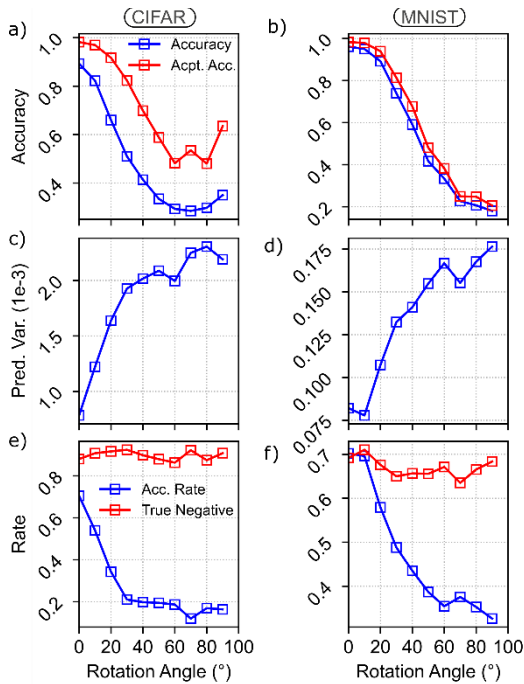


Fig. 7: Impact of input rotation on model performance and uncertainty for CIFAR-10 (left) and MNIST (right). (a–b) Predictive accuracy versus rotation angle, and accepted accuracy, computed only over predictions retained after uncertainty-based filtering. (c–d) Predictive variance that rises under domain shift (rotation angle). (e–f) Acceptance rate (fraction of inputs retained after filtering) and true negatives (fraction of rejected samples that would otherwise have been misclassified if accepted).

and the strong anti-correlation between accuracy and variance highlights that our approach enables effective post hoc uncertainty quantification in a hardware-compatible and computationally efficient manner. Finally, Fig. 7e–f reports the acceptance rate (the fraction of inputs the filter retains) and the true negatives, which are rejected samples that would otherwise have been misclassified. These results demonstrate how uncertainty filtering discards error-prone predictions, trading off coverage for improved accuracy and robustness.

## VI. POWER ANALYSIS AND BENCHMARKING

We benchmarked the power and energy efficiency of our BH architecture against other Bayesian CIM solutions. The combined MRAM array, VCO, and ripple counter consume 19 uW per column due to their event-driven operation. Fig. 8 compares the throughput-per-watt (TOPS/W) of our system against prior probabilistic CIM accelerators. The operation throughput (OPS) is given as:

$$\text{OPS} = i \times j \times \frac{1}{t} \quad (8)$$

Where  $i$  and  $j$  stands for the number of rows (16) and columns (64) and  $t$  the sampling time of 15.9 ns. Our design achieves an energy efficiency of 53 TOPS/W (18.9 fJ/OP) for 3-bit perturbations and 110 TOPS/W for 2-bit perturbations (4-count maximum). While some recent Bayesian CIM accelerators report efficiencies exceeding 100 TOPS/W [17], most do not specify the number of inferences runs and typically rely on repeated stochastic sampling or hundreds of Monte Carlo passes [13], [17], [18], [19], [20]. By contrast, our approach requires just 25 inference-time runs, with a total overhead of only 454 fJ, making the energy cost lower than prior Bayesian CIM accelerators. This extremely low cost enables uncertainty-aware inference without compromising

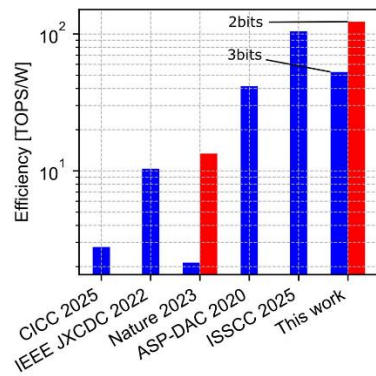


Fig. 8: Efficiency (TOPS/W) of Bayesian-based CIM solutions compared to the Bayesian Head (BH) [13], [17], [18], [19], [20]. Our efficiency is 53 TOPS/W for a 3-bit perturbation BH and 110 TOPS/W for 2-bit perturbation.

the baseline energy efficiency, making the proposed architecture especially well-suited for edge intelligence applications where both accuracy and confidence metrics are essential. Many prior Bayesian CIM works prioritize accuracy or model complexity at the expense of energy, area, or inference runs; our comparison focuses on uncertainty-aware efficiency and run count for edge deployment.

## VII. CONCLUSIONS

In this work, we presented a low-power, mixed-signal CIM architecture combining a deterministic Binary Neural Network with a Bayesian Head based on multi-pillar SOT-MRAM arrays. By introducing tunable stochastic perturbations to the BNN’s partial sums, the Bayesian Head enables uncertainty-aware inference with minimal hardware overhead and 19 uW power consumption.

The integration of a compact VCO-based ADC and a novel MP SOT-MRAM by-pass dropout strategy enables flexible Bayesian-inspired approximations. Unlike classical Bayesian methods that embed uncertainty during training, our approach applies calibrated perturbations post-training at inference, aligning with post hoc strategies such as dropout injection. Circuit-level simulations and algorithmic evaluations demonstrate competitive classification accuracy on CIFAR-10 and MNIST, while providing meaningful uncertainty estimation under domain-shift conditions. Crucially, our method requires only 25 inference-time runs, with an energy overhead of just 454 fJ, whereas most prior Bayesian CIM designs rely on repeated stochastic sampling or hundreds of Monte Carlo passes.

Importantly, the system achieves energy efficiency of 53 TOPS/W (18.9 fJ/OP) and 110 TOPS/W for for 3 and 2-bit perturbations, making it well-suited for edge intelligence scenarios where both performance and reliability are critical. The present evaluation focuses on lightweight BNN models to emphasize uncertainty-aware inference efficiency; extending the BH to more complex datasets is left for future work. These results establish our BH as a promising solution for enabling low-power, uncertainty-aware AI at the edge.

## ACKNOWLEDGMENT

This work was supported by internal CEA project PAST3C, the France 2030 government investment plan, managed by the French National Research Agency under grant reference PEPR Electronique (ANR-22-PEEL-0009); Horizon Europe grant 101182279 in the frame of Chips JU FAMES Pilot Line.

## REFERENCES

- [1] H. M. D. Kabir, A. Khosravi, M. A. Hosen, and S. Nahavandi, "Neural Network-Based Uncertainty Quantification: A Survey of Methodologies and Applications," *IEEE Access*, vol. 6, pp. 36218–36234, 2018, doi: 10.1109/ACCESS.2018.2836917.
- [2] D. Bonnet *et al.*, "Bringing uncertainty quantification to the extreme-edge with memristor-based Bayesian neural networks," *Nat Commun*, vol. 14, no. 1, p. 7530, Nov. 2023, doi: 10.1038/s41467-023-43317-9.
- [3] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Appendix," May 2016.
- [4] C. Blundell, J. Cornebise, K. Kavukcuoglu, W. Com, and G. Deepmind, "Weight Uncertainty in Neural Networks Daan Wierstra."
- [5] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," Dec. 2022, [Online]. Available: <http://arxiv.org/abs/1312.6114>
- [6] S. Liu *et al.*, "Bayesian neural networks using magnetic tunnel junction-based probabilistic in-memory computing," *Frontiers in Nanotechnology*, vol. 4, Oct. 2022, doi: 10.3389/fnano.2022.1021943.
- [7] A. Loquercio, M. Segu, and D. Scaramuzza, "A General Framework for Uncertainty Estimation in Deep Learning," *IEEE Robot Autom Lett*, vol. 5, no. 2, pp. 3153–3160, Apr. 2020, doi: 10.1109/LRA.2020.2974682.
- [8] E. Ledda, G. Fumera, and F. Roli, "Dropout injection at test time for post hoc uncertainty quantification in neural networks," *Inf Sci (N Y)*, vol. 645, p. 119356, Oct. 2023, doi: 10.1016/j.ins.2023.119356.
- [9] J. Doevenspeck *et al.*, "SOT-MRAM Based Analog in-Memory Computing for DNN Inference," in *2020 IEEE Symposium on VLSI Technology*, IEEE, Jun. 2020, pp. 1–2. doi: 10.1109/VLSITechnology18217.2020.9265099.
- [10] L. Geiger and P. Team, "Larq: An Open-Source Library for Training Binarized Neural Networks," *J Open Source Softw*, vol. 5, no. 45, p. 1746, Jan. 2020, doi: 10.21105/joss.01746.
- [11] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1," Mar. 2016.
- [12] S. T. Ahmed, K. Danouchi, M. Hefenbrock, G. Prenat, L. Anghel, and M. B. Tahoori, "Scale-Dropout: Estimating Uncertainty in Deep Neural Networks Using Stochastic Scale," Jan. 2024.
- [13] A. Lu, Y. Luo, and S. Yu, "An Algorithm-Hardware Co-Design for Bayesian Neural Network Utilizing SOT-MRAM's Inherent Stochasticity," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 8, no. 1, pp. 27–34, Jun. 2022, doi: 10.1109/JXCDC.2022.3177588.
- [14] K. Danouchi, G. Prenat, and L. Anghel, "Spin Orbit Torque-based Crossbar Array for Error Resilient Binary Convolutional Neural Network," in *2022 IEEE 23rd Latin American Test Symposium (LATS)*, IEEE, Sep. 2022, pp. 1–6. doi: 10.1109/LATS57337.2022.9936951.
- [15] P. K. AMIRI and K. L. WANG, "VOLTAGE-CONTROLLED MAGNETIC ANISOTROPY IN SPINTRONIC DEVICES," *SPIN*, vol. 02, no. 03, p. 1240002, Sep. 2012, doi: 10.1142/S2010324712400024.
- [16] T. Gokmen, M. Onen, and W. Haensch, "Training Deep Convolutional Neural Networks with Resistive Cross-Point Devices," *Front Neurosci*, vol. 11, Oct. 2017, doi: 10.3389/fnins.2017.00538.
- [17] D.-Q. You *et al.*, "14.1 A 22nm 104.5TOPS/W  $\mu$ -NMC- $\Delta$ -IMC Heterogeneous STT-MRAM CIM Macro for Noise-Tolerant Bayesian Neural Networks," in *2025 IEEE International Solid-State Circuits Conference (ISSCC)*, IEEE, Feb. 2025, pp. 1–3. doi: 10.1109/ISSCC49661.2025.10904540.
- [18] Z. M. Enciso *et al.*, "A 65 nm Bayesian Neural Network Accelerator with 360 fJ/Sample In-Word GRNG for AI Uncertainty Estimation," Jan. 2025.
- [19] K.-E. Harabi *et al.*, "A memristor-based Bayesian machine," *Nat Electron*, Dec. 2022, doi: 10.1038/s41928-022-00886-9.
- [20] H. Fan, M. Ferianc, Z. Que, X. Niu, M. Rodrigues, and W. Luk, "Accelerating Bayesian Neural Networks via Algorithmic and Hardware Optimizations," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 12, pp. 3387–3399, Dec. 2022, doi: 10.1109/TPDS.2022.3153682.