

Reliability and Assessment of an MRAM-based Non-Volatile Precharge-Free Ternary CAM for Real-Time Edge Inference

Oliver Caisaluisa^{1,2}, Fabian Seiler³, Luis Miguel Prócel¹, Marco Lanuzza^{4,5}, Adam Quotb^{6,2},
Nima TaheriNejad³, and Esteban Garzón^{4,5}

¹ Institute of Micro and Nanoelectronics (IMNE), Universidad San Francisco de Quito, Quito, Ecuador

² Université de Toulouse, Institut National Polytechnique (INP), Toulouse, France

³ Institute of Computer Engineering (ZITI), Heidelberg University, Germany

⁴ DIMES, University of Calabria, Rende, Italy. | ⁵ ICAR-CNR, Rende, Italy

⁶ LAAS-CNRS, Toulouse, France

Abstract—Non-volatile content-addressable memories (NV-CAMs) based on magnetic tunnel junctions (MTJs) are emerging as a compelling processing-using-memory (PuM) scheme for energy-efficient, massively-parallel pattern matching in AI inference and edge-computing workloads. This paper presents a comprehensive reliability and application-level study of a hybrid MTJ-CMOS NV precharge-free ternary CAM (NV-PTCAM) designed in a commercial 65 nm CMOS process. At the circuit level, we evaluate the impact of the number of ternary ‘don’t-care’ (X) states stored within the NV-TPCAM, and the impact of row segmentation. NV-PTCAM achieves a search delay below 0.25 ns, a search energy consumption of 1.82 fJ/cell, and a search error rate of about 0.02%, outperforming conventional NV-NOR and NV-NAND counterparts by up to 82% in search energy and up to 11.5× in search delay, on average. We apply an endurance model to the NV-PTCAM search workload and demonstrate that the number of cycles to failure exceeds 10^{13} for pulse amplitudes as high as 0.656 V and pulse widths in the range 100–500 ps, establishing the practical reliability of the NV-PTCAM for explainable artificial intelligence (xAI) models useful in wearable or edge devices. At the application level, we demonstrate the NV-PTCAM as the PuM inference engine of a real-time photoplethysmography (PPG)-based tachycardia classifier. The NV-PTCAM-based Random-Forest-based PuM system achieves a mean accuracy of 94.8% and an F1-score of 86.7% on the BIDMC clinical dataset, while the low SER of the NV-PTCAM has a negligible impact on classification performance.

Index Terms—Content-Addressable Memory, Ternary CAM, NV-CAM, MRAM, DMTJ, Precharge-Free, Segmentation, Reliability, PPG, Tachycardia, Edge Inference, Decision Tree, Random Forest.

I. INTRODUCTION

Content-addressable memories (CAMs) are associative memories that perform massively parallel comparisons between a query pattern and every stored word in a single clock cycle [1]. By returning the address (or addresses) of matching entries in $\mathcal{O}(1)$ time regardless of array size, CAMs unlock

throughput that sequential search cannot match. Their applications span high-speed routing [2], fully-associative caches [3], [4], data analytics [5]–[7], and machine-learning applications [8]–[10].

Conventional CMOS-based CAMs, however, suffer from large static power and large area footprint. Non-volatile CAMs (NV-CAMs) address this by using emerging non-volatile memory (eNVM) devices, including resistive RAM (RRAM), phase-change memory (PCM), and magnetic RAM (MRAM), storing NV information useful for battery-powered wearable devices [11]. Among eNVM technologies, MRAM-based on MTJs stands out for its CMOS compatibility, fast switching, and practically unlimited write endurance. The introduction of double-barrier MTJs (DMTJs) further reduces the critical switching current, enabling energy-efficient write operations while preserving low-power search operations in NV-CAM designs [12], [13].

Building on this foundation, the NV-PTCAM [13] is the first non-volatile precharge-free CAM based on DMTJ devices. By eliminating the precharge phase entirely, NV-PTCAM breaks the fundamental energy overhead of conventional NOR-type NV-CAMs while also reducing the search error rates to about 0.02%, mainly due to the use of memory device redundancy, i.e., two devices to store one bit of information, thereby, increasing the search sensing margin. However, relatively high search voltages are required to maintain performant and reliable search operations, and their impact on practical applications has not yet been evaluated.

The use of TCAM-based decision trees has gained keen attention as a high-performance associative memory solution [14], [15]. At the same time, decision trees offer a compelling approach for efficient real-time edge classification in healthcare [16]–[18]. Photoplethysmography (PPG) sensors are ubiquitous in wearable devices and provide a non-invasive proxy of the cardiovascular state [19]. Detecting tachycardia (heart rate > 100 bpm) from PPG signals in real time, on-device, and at sub-milliwatt power is a clinically relevant

This work was supported by the Italian Ministry of University and Research (MUR) under grant number SOE_20240000022, the German Research Foundation (DFG) under Project 550835853, and Hector Stiftung.

target that conventional digital processors address only with significant energy overhead [20].

This paper presents a real-time edge inference accelerator using MRAM-based NV precharge-free ternary CAM (NV-PTCAM) for full decision-tree lookup in a single search cycle while ensuring energy-efficient operations. Specifically, the NV-PTCAM is designed using DMTJ technology and a commercial 65 nm process, extending the analysis and evaluation presented in [13]. A comprehensive search performance, reliability, and endurance evaluation are carried out at the circuit-level. At the application level, the NV-PTCAM is used as the processing-using-memory (PuM) inference architecture for real-time photoplethysmography (PPG)-based tachycardia classifier operating on the Beth Israel Deaconess Medical Center (BIDMC) clinical dataset, validating the end-to-end suitability of the architecture for always-on cardiac monitoring at the edge.

This paper makes the following contributions:

- A comprehensive circuit-level characterization of the NV-PTCAM, analyzing the impact of don't-care 'X' states and ML segmentation, outperforming the conventional ternary NV-NOR and NV-NAND counterparts by up to 82% in search energy and up to $11.5\times$ in search delay, on average, while achieving the lowest SER among all evaluated schemes.
- A reliability analysis of the MTJ endurance under realistic NV-PTCAM search workloads, confirming cycles-to-failure in the range 10^{13} – 10^{19} across the full operating window, establishing effectively unlimited search endurance for edge inference deployment.
- The first NV-PTCAM-based PuM inference engine for real-time PPG tachycardia classification, achieving 94.8% (92.3%) mean (balanced) accuracy with a random forest of 17 trees on the BIDMC clinical dataset, using only ~ 510 words of TCAM storage in a 0.107 mm^2 array footprint.

The DMTJ endurance analysis confirms that the number of cycles-to-failure reaches 10^{14} under the worst-case NV-PTCAM search conditions, establishing an effectively unlimited operational lifetime that far exceeds the 10^{12} -cycle write endurance of commercial STT-MRAM.

II. BACKGROUND

A. NV-PTCAM

Fig. 1 shows the NV-PTCAM proposed in [13]. It is a voltage-divider-based NV-CAM cell that stores one ternary bit across two series-stacked pairs of DMTJs, for a total of four devices per cell. Each bit is encoded in the relative resistance configuration of the left ($Jl_{eq} = Jl_1 + Jl_2$) and right ($Jr_{eq} = Jr_1 + Jr_2$) DMTJ stacks: logic '0' ('1') places the left (right) stack in the high-resistance state (HRS) and the right (left) stack in the low-resistance state (LRS), while the 'don't-care' (X) state places all four devices in HRS. Stacking two DMTJs per branch doubles the absolute resistance HRS/LRS ratio $\Delta R_{\text{stack}} = 2\Delta R$ compared to single-device schemes,

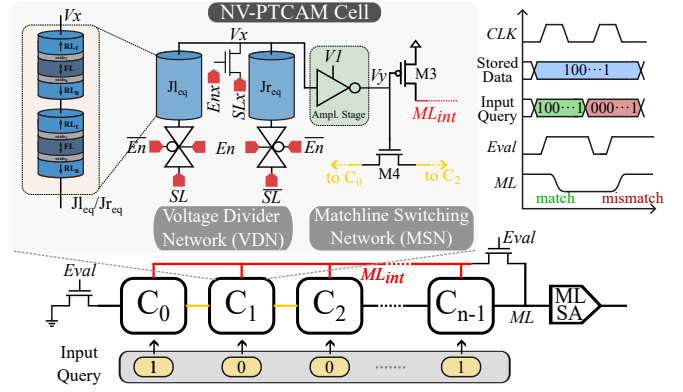


Fig. 1. Non-volatile precharge-free content-addressable memory [13].

TABLE I
SEARCH OPERATIONS OF NV-PTCAM CELLS.

	Control signals		Search signals		Stored	ML
Operation	En/En	Enx	SL/SL	SLx	data	result
Search ‘1’	$V_{DD}/0$	0	$V_s/0$	0	0/1/X	Mis./Match/Match
Search ‘0’	$V_{DD}/0$	0	$0/V_s$	0	0/1/X	Match/Mis./Match
Search ‘X’	$V_{DD}/0$	0	0/0	0	any	Match
Note: Match/Mismatch means ML at $ground/V_{DD}$						

Note: Match/Mismatch means ML at ground/ V_{DD}

which directly enlarges the voltage sensing margin (ΔV_x) available to the matchline switching network (MSN) in Fig. 1.

Writing a logic '0' or '1' is accomplished by applying a bidirectional current pulse through the DMTJ stacks, exploiting spin-transfer-torque (STT) switching to set each stack to the target HRS or LRS state. Specifically, the write signal is sent from the search lines (SL , $\overline{SL}$, SLx) while asserting the enable signals (En , $\overline{En}$, and Enx), as per [13]. During a search operation ($Eval = 1$), input circuitry drives a voltage pulse V_s from the SL or $\overline{SL}$ depending on the search operation, as detailed in Table I. This applied V_s across the DMTJ stack results in a midpoint voltage (V_x), which is translated to full-rail logic levels through the amplification stage ($V1 = 1.25V$) and further evaluated by the complementary pull-up and pull-down transistors $M3$ and $M4$ in the MSN: a full-row match ($V_y = 1$ in Fig. 1) forms a discharge path through all cells, while any mismatch ($V_y = 0$ in Fig. 1) charges the internal matchline (ML_{int}), all without requiring an explicit precharge phase. The ML sense amplifier (MLSA), one per row, signals the output of the search operation. To prevent unintended STT switching of the DMTJ stacks during search, the pulse width is kept within a safe timeframe [13], with a hold time before sending the next pulse.

B. Decision Trees & Healthcare Applications

Decision Trees (DT) map naturally onto the TCAM inference paradigm: each internal node evaluates a threshold condition on an input feature, which directly corresponds to the lookup operation performed by a TCAM row [14], [15]. By encoding all threshold conditions of multiple trees into TCAM arrays, a Random Forests (RF) classifier can be built, enabling the entire inference step to be carried out in a single clock cycle. This gives $\mathcal{O}(1)$ classification latency regardless of tree depth, compared to $\mathcal{O}(D)$ for the sequential counterpart. Beyond

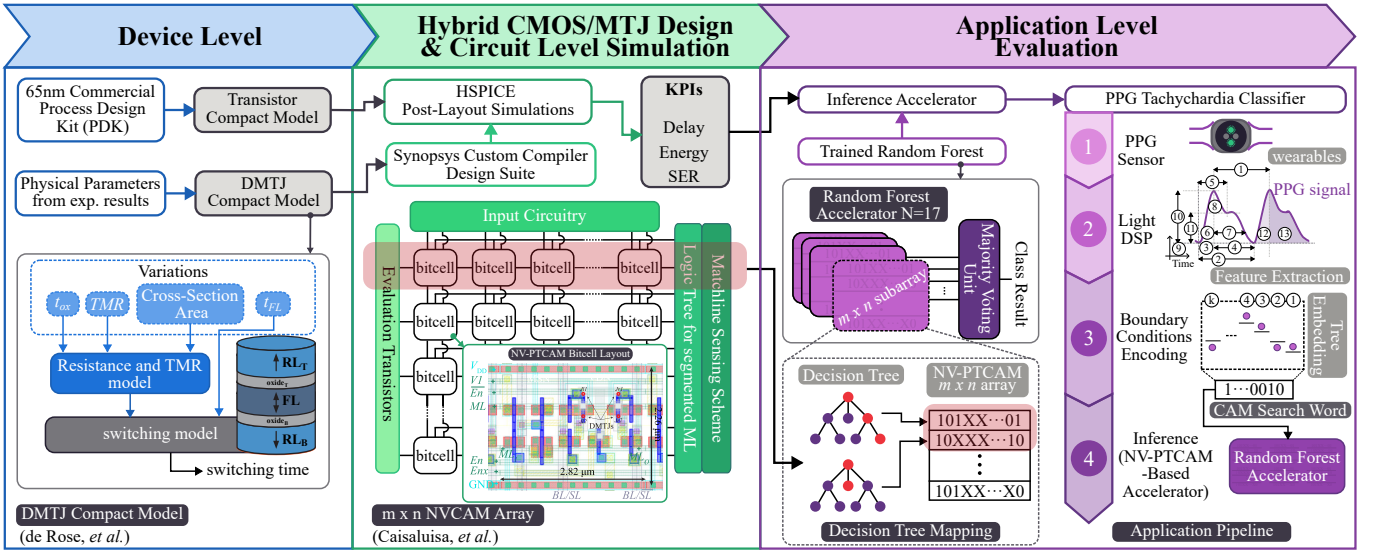


Fig. 2. The circuit-to-application evaluation framework used in this work. (Left) Compact models for hybrid CMOS-MTJ simulations. (Center) NV-PTCAM $m \times n$ cell array that maps a single decision tree (30 words of 32 bits). (Right) Inference accelerator composed of $N = 17$ parallel NV-PTCAM sub-arrays with majority-voting unit. End-to-end pipeline: ① PPG sensor acquisition, ② lightweight DSP feature extraction (13 morphological features), ③ boundary-condition encoding into a 32-bit ternary search vector, ④ parallel TCAM search with $\mathcal{O}(1)$ classification output. De Rose, *et al.* [21]; Caisaluisa, *et al.* [13].

their hardware compatibility, DT models are attractive for wearable healthcare applications due to their low computational complexity, compact memory footprint, and suitability for embedded inference [16]–[18], [22]. In particular, PPG-based cardiovascular monitoring is well suited for lightweight edge processing, since physiological features such as pulse rate, interval variability, and waveform morphology can be efficiently classified using threshold-based models [19], [20], [23]–[25]. These characteristics make DT inference and NV-PTCAM technology an ultra-low-power PuM architecture for wearable sensing applications [26].

III. EVALUATION FRAMEWORK

Fig. 2 illustrates the complete evaluation framework used in this work, spanning from the NV-PTCAM circuit level up to the real-time inference accelerator, both detailed hereafter.

A. Circuit-Level Evaluation Flow

All Circuit-Level (refer to device-to-circuit illustration in Fig. 2) are obtained through post-layout hybrid CMOS-DMTJ simulations in Synopsys Custom Compiler using a commercial 65 nm PDK and a Verilog-A DMTJ compact model [21]. The reported figures-of-merit are based on 3000 Monte Carlo samples, and correspond to a nominal supply voltage of $V_{DD} = 1.1$ V, a 3σ corner unless stated otherwise. To cover different ternary workloads scenarios, we evaluate the main figures-of-merit (search delay, search energy, and SER) under different X-state ratios stored within the NV-PTCAM.

From Fig. 2, the NV-PTCAM is organized as an $m \times n$ array of ternary CAM cells, where each row stores one 32 bit ternary word. Word length is also extended beyond the single-segment through row segmentation: the 32-bit row is partitioned into $N_s = 4$ segments of 8 bits each, whose local MLs are combined by a logic tree to produce the global match or mismatch signal.

B. Application-Level Evaluation: PPG Tachycardia Classifier

From Fig. 2, at the application-level evaluation we use NV-PTCAM as a DT to accelerate RF inference. A single NV-PTCAM 30×32 cell array uses 30 rows to encode one complete DT: each row stores a root-to-leaf path as a ternary word, with ‘X’ masking thresholds that are irrelevant on that path, as highlighted in red in Fig. 2. A single search against the an NV-PTCAM array evaluates all stored paths simultaneously, returning the matching leaf label in $\mathcal{O}(1)$ time regardless of tree depth.

The full system pipeline, shown in Fig. 2, proceeds in four stages: ① A PPG sensor from a wearable device acquires the raw waveform continuously. ② A lightweight DSP block extracts 13 morphological features per pulse, including pulse period, systolic and diastolic peak amplitudes and timings, dicrotic notch position, and several derived area and width measures, entirely in the time domain. ③ Each feature is compared against the k learned boundary conditions of the trained tree ensemble, and the resulting binary decisions are assembled into a 32-bit ternary search vector. ④ The query vector is the input of the proposed RF accelerator. This query is broadcast simultaneously to all 17 NV-PTCAM sub-arrays; the matching row in each sub-array returns a leaf label, and majority voting over the N labels yields the final tachycardia or normal classification in $\mathcal{O}(1)$ time.

1) Dataset and Task Definition

We evaluate the NV-PTCAM-based RF classifier on the Beth Israel Deaconess Medical Center (BIDMC) PPG dataset [27], which contains recordings from 53 patients, yielding approximately 3000 samples. The task is binary classification: tachycardia (heart rate > 100 bpm) vs. normal, with a class imbalance of 17% tachycardia vs. 83% normal. Specifically, we use 10 random patient-wise 80 (20) train (test) splits with non-overlapping train and test windows. Each split

contains both classes and is evaluated using 1000 inference runs per sample.

2) Decision-Tree Embedding in NV-PTCAM

Each path from root to leaf in a trained DT is encoded as one TCAM word. A feature value below (above) a threshold sets the corresponding TCAM bit to '0' ('1'); thresholds not present along a root to leaf path are masked with 'X'. The matching rows present the stored label field and returns the classification decision. A single NV-PTCAM array search thus evaluates the entire DT in one shot. For a RF of $N = 17$ trees, each tree is mapped to a NV-PTCAM sub-array. During inference, all sub-arrays operate in parallel, and majority voting over the N results produces the final prediction.

3) Robustness & Reliability Assessment

Two analyses are carried out to evaluate the practical robustness and long-term reliability of the NV-PTCAM-based RF inference accelerator. The first one is the SER impact on classification. During each search operation, the NV-PTCAM is subject to search errors arising from the limited sensing margin of the matchline switching network. To quantify the impact of SER on the RF classifier, we inject errors at the output of each individual NV-PTCAM sub-array (i.e., at the DT level) with rates equal to the SER values obtained from post-layout Monte Carlo simulations at the circuit level. We evaluate the impact of SER on accuracy, balanced accuracy, precision, recall, and F1-score over the full test set.

The second analysis is the endurance under voltage stress. Each NV-PTCAM search cycle applies a voltage pulse V_s across the DMTJ stacks, leading to constant stress that may degrade the MTJ tunnel barrier over time. To assess the operational lifetime of the NV-PTCAM under realistic inference workloads, we apply the semi-empirical endurance model in [28], which estimates the number of cycles to failure N_C as:

$$N_C = N_{C0} \frac{n_{D0}}{n_D}; n_D = n_{D0} \left[\frac{t_+}{t_0} e^{\alpha V_+} + e^{\beta |V_-|} + \frac{t_-}{t_0} e^{\alpha |V_-|} e^{\beta V_+} \right], \quad (1)$$

where V_+ (V_-) are the positive (negative) pulse amplitudes, t_+ and t_- are the corresponding pulse widths, and $\alpha, \beta, N_{C0}, n_{D0}, t_0$ are technology-dependent fitting parameters [28]. We apply symmetric pulse conditions ($|V_+| = |V_-|$) with $t_p \in [100 \text{ ps}, 500 \text{ ps}]$, consistent with the safe-timeframe search pulse of the NV-PTCAM, and sweep V_p across the range relevant to our circuit operating point.

IV. RESULTS

A. Circuit-Level: Energy, Delay, SER

Fig. 3 reports the search time (t_{search}), search energy per row ($E_{\text{search,row}}$), and SER of NV-PTCAM (segmented and non-segmented), NV-NOR, and NV-NAND as a function of the percentage of 'X' states (25%, 50%, 75%) stored within a word.

The segmented NV-PTCAM presents the best figures-of-merit all the cases. Its t_{search} remains at about 0.2 ns for all X-state stored cases, in contrast to NV-NOR (≈ 1.79 ns) and NV-NAND (increasing from 4.1 ns at 25% to 6.2 ns at 75%), which both suffer due to their precharge phase during a search

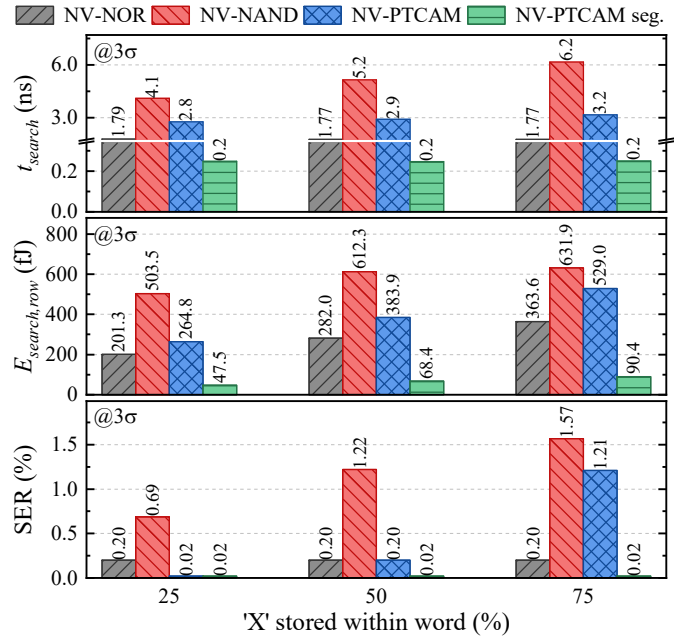


Fig. 3. Ternary X-state ratio impact on (top) search delay t_{search} , (center) search energy per row $E_{\text{search,row}}$, and (bottom) SER, for NV-NOR, NV-NAND, NV-PTCAM (non-segmented), and NV-PTCAM (4-segment) at 3σ . All results from 3000 Monte Carlo samples.

operation. The non-segmented NV-PTCAM rises from 2.8 ns to 3.2 ns as the number of X-states stored increases. This is mainly due to the longer matchline discharge path when more cells present a weak pull-down through the 'X' configuration [13].

The $E_{\text{search,row}}$ of the segmented NV-PTCAM ranges from 47.5 fJ at 25% 'X' to 90.4 fJ at 75% X, remaining well below NV-NOR (201–364 fJ) and NV-NAND (503–632 fJ) across the entire range. The energy increase of the segmented NV-PTCAM, NV-NOR, and NV-NAND schemes with higher X-state ratio is primarily due to the increased hold-pulse duration required to avoid overwrite hazards, ensuring reliable ML evaluation [13].

The SER of the segmented NV-PTCAM holds at 0.02% across all X-state ratios, the lowest among all evaluated schemes. NV-NOR maintains 0.20%, while NV-NAND degrades from 0.69% at 25% 'X' to 1.57% at 75% 'X', a $2.3\times$ increase driven by charge-sharing effects in the NAND chain that are exacerbated by the partial conductance of X-encoded cells. The non-segmented NV-PTCAM shows a moderate increase from 0.02% to 1.21%.

B. Circuit-Level: Matchline Segmentation

Fig. 4 shows the impact of the number of 8-bit segments (N_s) on t_{search} and $E_{\text{search,bit}}$ for the NV-PTCAM at a fixed 32-bit word. The t_{search} decreases monotonically with N_s , decreasing (about 99%) from 2.80 ns at $N_s = 1$ to 0.03 ns at $N_s = 16$. This speedup is mainly due to the reduction in ML discharge path length per segment.

$E_{\text{search,bit}}$ is significantly higher for low numbers of segments ($N_s = 1$ and $N_s = 2$). This is mainly because the corresponding t_{search} exceeds the 500 ps safe search pulse

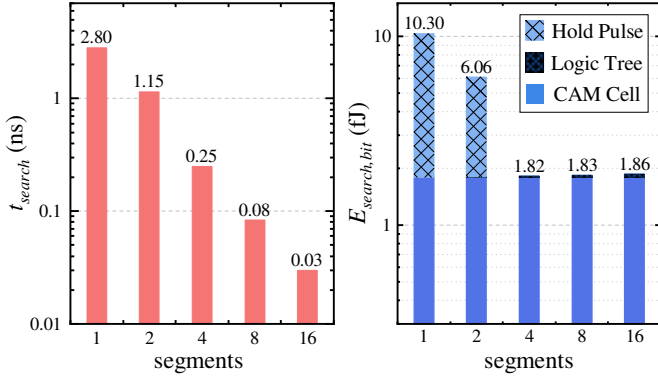


Fig. 4. Search delay (left) and search energy per bit (right) as a function of the number of 8-bit segments $N_s \in \{1, 2, 4, 8, 16\}$ for NV-PTCAM (32-bit word). Energy is breakdown accounts for CAM-cell energy, logic-tree switching energy, and hold-pulse energy.

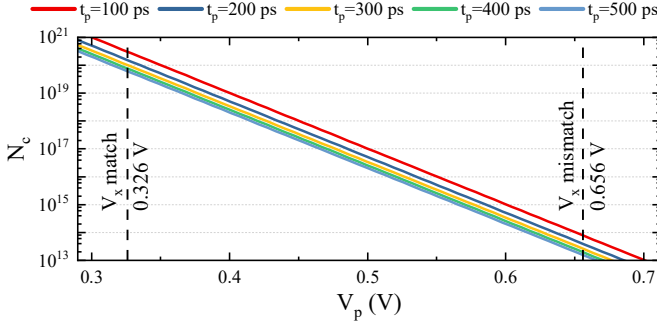


Fig. 5. Number of cycles to failure N_C vs. pulse amplitude V_p for symmetric search pulses with width $t_p \in \{100, 200, 300, 400, 500\}$ ps, using the Carboni endurance model applied to the NV-PTCAM search operation.

window, requiring a hold time between consecutive pulses to avoid overwrite hazards, as detailed in [13]. As the number of segments increases ($N_s \geq 4$), t_{search} falls below 500 ps, eliminating the need for inter-pulse hold time. Therefore, the $E_{search,bit}$ remains roughly the same (between 1.82-1.86 fJ/bit), with only a minor overhead due to the additional logic-tree circuitry of the segmented NV-PTCAM word.

C. Circuit-Level: DMTJ Endurance under Search Operations

Fig. 5 presents the endurance analysis of the DMTJ stack under the NV-PTCAM search operation, using the Carboni endurance model [28] evaluated over the full operating window defined by the match and mismatch voltage levels. Specifically, these voltage levels are obtained from the voltage-sharing mechanism of the series-stacked DMTJ configuration, where the applied search voltage is divided across both devices in each brach (refer to Fig. 1).

At the match boundary ($V_p = 0.326$ V), N_C exceeds 10^{19} for all evaluated search pulse widths, confirming effectively unlimited endurance under match conditions. At the mismatch boundary ($V_p = 0.656$ V), N_C remains in the range 10^{13} - 10^{14} , with the most aggressive condition ($t_p = 500$ ps) yielding $N_C \approx 10^{14}$. Even this worst-case scenario is several orders of magnitude above the 10^{12} write endurance of commercial STT-MRAM, and far beyond any realistic search workload. These results validate the safe-timeframe methodology and provide a quantitative reliability specification.

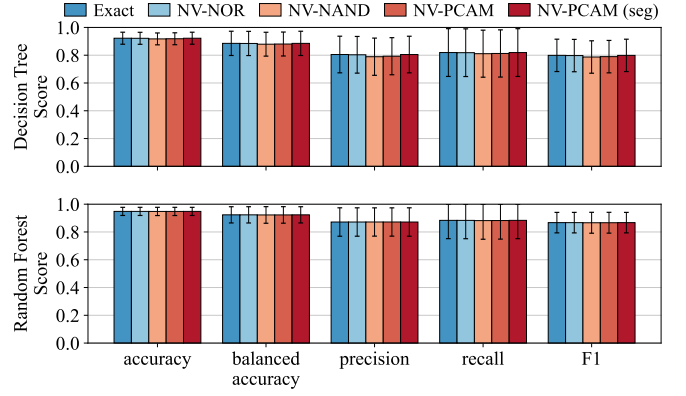


Fig. 6. Classification metrics (accuracy, balanced accuracy, precision, recall, F1-score) for the (top) single Decision Tree and (bottom) Random Forest ($N = 17$) as a function of NV-PTCAM SER. Error bars indicate $n = 1000$ randomized repetitions per inference over 10 train or test splits.

D. Application-Level Results: PPG Tachycardia Classification

1) Decision Tree

Fig. 6 shows the classification results. A single decision tree achieves a mean (balanced) accuracy of 92.2% (88.5%) at zero search errors. Precision, recall, and F1-score are approximately 80.5%, 81.9%, and 79.9%, respectively. The NV-PTCAM (segmented) presents a SER of 0.02% comparable to the exact baseline across all metrics. Overall, performance is stable across all NV-CAM designs, confirming that the search error of the NV-CAM designs have a negligible impact on single-tree classification.

2) Random Forest ($N = 17$)

From Fig. 6, ensembling 17 DTs with majority voting lifts mean accuracy to 94.8% and balanced accuracy to 92.3%, with precision, recall, and F1-score reaching approximately 87.2%, 88.4%, and 86.7%, respectively. The RF is immune to the obtained SERs at the circuit-level evaluation. This robustness arises mainly from the redundancy of majority voting.

3) Memory Footprint and Performance

Each decision tree is stored in 30 words of 32 bits (960 cells) NV-PTCAM sub-array. Considering our cell area footprint of about $6.37 \mu\text{m}^2$. The array area is about $6115 \mu\text{m}^2$ including the periphery overhead ($\sim 3.1\%$) of $189.6 \mu\text{m}^2$ due to evaluation transistors and logic tree circuitry for ML segmentation. Each DT real-time edge inference (tachycardia in this work) executes in a single $t_{search} = 0.25$ ns search cycle, consuming $E_{search} = 960 \times 1.82 = 1747$ fJ (≈ 1.75 pJ) per inference and yielding a throughput of $1/t_{search} = 4.35$ GInferences/s.

For the random forest of $N = 17$ trees, the 17 NV-PTCAM sub-arrays operate in parallel, so the search latency is unchanged at $t_{search} = 0.25$ ns while the total storage grows to $17 \times 30 = 510$ words ($\approx 0.107 \text{ mm}^2$). The 17 leaf labels of the DTs are combined by a NAND-based majority-vote logic tree [29] comprising of 10 three-input NAND voters (38 ps, $27 \mu\text{W}$, $54 \mu\text{m}^2$ each at 65 nm), adding at most ~ 150 ps, ~ 40 fJ, and $< 0.52\%$ area overhead. Global match detection across the 30 stored rows of a sub-array adds a few tens of ps. The total end-to-end inference latency is about 0.38 ns (2.63 GInferences/s) with a per-inference energy of

about 29.7 pJ. Note that extra energy and latency overhead may be added depending on the MLSA circuitry.

From a reliability standpoint, the N_C values reported in Section IV-C translate directly into an operational lifetime for the inference accelerator. Although the NV-PTCAM is capable of about 2.63 GInferences/s, the PPG tachycardia classification task requires only one inference per cardiac pulse (~ 1 -3 Hz), since no additional diagnostic information is gained by classifying faster than the physiological event rate. Under nominal conditions ($V_p = 0.326$ V, $N_C > 10^{18}$), the search endurance is effectively unlimited, exceeding the typical 10^{12} -cycle write endurance of commercial STT-MRAM, confirming that the NV-PTCAM search operation imposes no practical reliability constraint on the inference accelerator

V. COMPARISON

Table II summarizes the figures-of-merit of all four NV-CAM schemes. The segmented NV-PTCAM achieves the best combination of search energy, search delay, and SER. Its t_{search} (0.25 ns) is 7.2 $\times$, 15.8 $\times$, and 11.6 $\times$ faster than NV-NOR, NV-NAND, and the non-segmented NV-PTCAM, respectively, while its E_{search} (1.82 fJ/cell) is 75.8%, 87.2%, and 82.3% lower. The reduction in search delay is primarily enabled by the 4-segment ML architecture, which shortens the discharge path and eliminates inter-pulse delays by keeping t_{search} within the safe timeframe window. Indeed, NV-NOR requires 4 search pulses per operation and NV-NAND requires 8, introducing significant pulse-overhead latency, whereas both NV-PTCAM configurations require only a single pulse.

At the inference accelerator level, the circuit-level (i.e., array-level) gains add up to 1.75 pJ (29.7 pJ) per DT (RF) inference, respectively, compared to 7.24 pJ (123 pJ) for NV-NOR, 13.6 pJ (231 pJ) for NV-NAND, and 9.89 pJ (168 pJ) for the non-segmented NV-PTCAM. This is a reduction of up to 7.8 $\times$ at the RF level for the proposed NV-PTCAM inference accelerator. The 16% smaller cell area of NV-PTCAM ($6.37 \mu\text{m}^2$ vs. $7.59 \mu\text{m}^2$) further reduces the total RF array footprint to 0.107 mm^2 , compared to 0.124 - 0.126 mm^2 for the precharge-based schemes. Note that, at the memory array level, the logic-tree overhead of the 4-segment architecture adds a $\sim 2.5\%$ area penalty over the non-segmented NV-PTCAM, yielding a DT array footprint of $6304 \mu\text{m}^2$ and an RF array footprint of 0.107 mm^2 , both still well below the 7286 - $7440 \mu\text{m}^2$ and 0.124 - 0.126 mm^2 of the precharge-based schemes.

Despite their higher SER values, all four schemes yield about the same RF classification performance, with 94.8% accuracy and F1-scores within 0.1% of each other, confirming that majority voting over $N = 17$ trees renders the inference accelerator robust to the search error rates of any evaluated NV-CAM design. The write performance and the number of DMTJs per cell are identical across all schemes, confirming that all architectural differences are confined to the search path.

Note that NV-PTCAM parallel search completes one RF inference in less than 1 ns, over 100 $\times$ faster than a sequential SRAM-based LUT implementation at the same 65 nm node,

TABLE II
PERFORMANCE COMPARISON OF ALL NV-CAM SCHEMES DESIGNED IN THIS WORK (TERNARY, 32-BIT WORD, 3σ , POST-LAYOUT).

Parameter	NV-NOR	NV-NAND	NV-PTCAM (no seg.)	NV-PTCAM (seg.)
Technology (nm)	65	65	65	65
MTJ type	DMTJ	DMTJ	DMTJ	DMTJ
Precharge	YES	YES	NO	NO
# segments	1	4	1	4
# search pulses	4	8	1	1
# transistors / cell	10	10	9	9
# MTJs / cell	4	4	4	4
Cell area (μm^2)	7.59	7.59	6.37	6.37
t_{search} (ns)	1.80	3.94	2.91	0.25
E_{search} (fJ/cell)	7.54	14.2	10.3	1.82
SER (%)	0.22	0.99	0.36	0.02
t_{write} / cell (ns)	5.5	5.5	5.5	5.5
E_{write} (fJ/cell)	187	187	187	187
Array area – DT (μm^2)	7286	7440	6150	6304
Array area – RF (mm^2)	0.124	0.126	0.105	0.107
E_{search} – DT (pJ)	7.24	13.6	9.89	1.75
E_{search} – RF (pJ)	123	231	168	29.7
t_{search} – DT/RF (ns)	1.80	3.94	2.91	~ 0.5
Accuracy – RF (%)	94.8	94.8	94.8	94.8
F1-score – RF (%)	86.7	86.6	86.6	86.7

which requires ≥ 28 ns ($N = 17$ trees $\times 1.67$ ns per access at 600 MHz [30]). While the SRAM baseline achieves lower per-inference energy (~ 6.4 pJ vs. 29.7 pJ), this comes at the cost of volatile storage and continuous leakage, whereas the NV-PTCAM retains the trained model with zero standby power, a key feature for always-on wearable inference.

VI. CONCLUSION

This paper presents a comprehensive reliability and application-level study of a NV precharge-free ternary CAM (NV-PTCAM), covering the impact of ternary X storage on search errors, word-level matchline segmentation, DMTJ endurance under search operations, and real-time edge inference for cardiac monitoring. At the circuit level, the 4-segment NV-PTCAM presents a search delay of 0.25 ns and an SER of 0.02%, outperforming NV-NOR and NV-NAND by up to 11.5 $\times$ in search delay and up to two orders of magnitude in SER. The DMTJ endurance analysis confirms that the number of cycles-to-failure reaches 10^{14} under the worst-case NV-PTCAM search conditions. At the application level, the NV-PTCAM-based PPG tachycardia classifier achieves a mean (balanced) accuracy of 94.8% (92.3%) with a random forest of 17 trees on the BIDMC clinical dataset, using only ~ 510 words of TCAM storage in 0.107 mm^2 . The low intrinsic SER of the NV-PTCAM has a negligible impact on classification performance, highlighting the architecture as a practical PuM inference engine for real-time, energy-efficient edge inference in wearable healthcare devices.

REFERENCES

- [1] T. Molom-Ochir, B. Taylor, H. Li, and Y. Chen, "Advancements in content-addressable memory (cam) circuits: State-of-the-art, applications, and future directions in the ai domain," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 72, no. 8, pp. 3971–3982, 2025.
- [2] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: a tutorial and survey," *IEEE J. of Solid-State Circuits*, vol. 41, no. 3, pp. 712–727, 2006.
- [3] E. Garzón, R. Hanhan, M. Lanuzza, A. Teman, and L. Yavits, "FASTA: Revisiting Fully Associative Memories in Computer Microarchitecture," *IEEE Access*, vol. 12, pp. 13 923–13 943, 2024.
- [4] S. Mishra, T. V. Mahendra, and A. Dandapat, "A 9-T 833-MHz 1.72-fJ/bit/search quasi-static ternary fully associative cache tag with selective matchline evaluation for wire speed applications," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 11, pp. 1910–1920, 2016.
- [5] E. Garzón, V. Galindo, Y. Harary, A. Teman, and L. Yavits, "Integrated BSI Bacteria Identifier-On-Chip Using Approximate k-mer Matching," *Scientific Reports*, pp. 1–15, 2026.
- [6] A. Shaban, T.-H. Hou, and M. Suri, "SOT-MRAM-based approximate content addressable memory for DNA classification," *IEEE Transactions on Electron Devices*, 2024.
- [7] Z. Jahshan, I. Merlin, E. Garzón, and L. Yavits, "CAM4: In-Memory Viral Pathogen Genome Classification using Similarity Search Dynamic Content-Addressable Memory," *Transactions on Emerging Topics in Computing*, vol. 13, no. 4, pp. 1341–1355, 2025.
- [8] Y.-J. Chang, K.-L. Tsai, and X.-Y. Chen, "Reconfigurable 9T static RAM design with dynamic switching between content-addressable memory and in-memory computing," *IEEE Transactions on Circuits and Systems for Artificial Intelligence*, 2025.
- [9] X. Yin, F. Müller, A. F. Laguna, C. Li, Q. Huang, Z. Shi, M. Lederer, N. Laleni, S. Deng, Z. Zhao *et al.*, "Deep random forest with ferroelectric analog content addressable memory," *Science advances*, vol. 10, no. 23, p. eadk8471, 2024.
- [10] Y. Harary, A. Sharoni, E. Garzón, M. Lanuzza, A. Teman, and L. Yavits, "PIC-BNN: A 128-kbit 65nm Processing-in-CAM-Based End-to-End Binary Neural Network Accelerator," in *2025 Cross-Disciplinary Conference on Memory-Centric Computing (CCMCC)*. IEEE, 2025, pp. 1–7.
- [11] K. Lee, J. J. Kan, and S. H. Kang, "Unified embedded non-volatile memory for emerging mobile markets," in *Proceedings of the 2014 international symposium on Low power electronics and design*, 2014, pp. 131–136.
- [12] C. Mosquera, E. Garzón, M. Lanuzza, A. Teman, and L. Yavits, "AM5: Bulk Logic-in-Memory Using MRAM NAND Crossbar," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 16, no. 2, pp. 203–217, 2026.
- [13] O. Caisaluisa, E. Garzón, E. Holguín, M. Lanuzza, and P. Luis Miguel, "NV-PCAM: Non-Volatile Precharge-Free Content-Addressable Memory," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 73, no. 3, pp. 1539–1552, 2026.
- [14] Y.-C. Liao, C.-L. Tsai, Y.-H. Chang, C. Slimani, J. Boukhobza, and T.-W. Kuo, "Retention: Resource-efficient tree-based ensemble model acceleration with content-addressable memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2026.
- [15] M. Rakka, M. E. Fouda, R. Kanj, and F. Kurdahi, "Dt2cam: A decision tree to content addressable memory framework," *IEEE Transactions on Emerging Topics in Computing*, vol. 11, no. 3, pp. 805–810, 2023.
- [16] E. F. Siddiqui, T. Ahmed, and S. K. Nayak, "A decision tree approach for enhancing real-time response in exigent healthcare unit using edge computing," *Measurement: Sensors*, vol. 32, p. 100979, 2024.
- [17] S. Alex, K. Dhanaraj, and P. Deepthi, "Private and energy-efficient decision tree-based disease detection for resource-constrained medical users in mobile healthcare network," *IEEE Access*, vol. 10, pp. 17 098–17 112, 2022.
- [18] I. Sudha, O. Gurrapu, P. Johri, J. Rakhimberganov, R. Kurbanioyozova, and N. V. Patel, "Designing Smart Healthcare Wearables for Remote Patient Monitoring Using Decision Tree-Based Anomaly Detection Algorithms," in *2025 International Conference on Networks and Cryptology (NETCRYPT)*. IEEE, 2025, pp. 887–892.
- [19] P. H. Charlton, P. A. Kyriacou, J. Mant, V. Marozas, P. Chowieńczyk, and J. Alastruey, "Wearable photoplethysmography for cardiovascular monitoring," *Proceedings of the IEEE*, vol. 110, no. 3, pp. 355–381, 2022.
- [20] M. Risso, A. Burrello, D. J. Pagliari, S. Benatti, E. Macii, L. Benini, and M. Pontino, "Robust and energy-efficient ppg-based heart-rate monitoring," in *2021 IEEE international symposium on circuits and systems (ISCAS)*. IEEE, 2021, pp. 1–5.
- [21] R. De Rose, M. d'Aquino, G. Finocchio, F. Crupi, M. Carpentieri, and M. Lanuzza, "Compact Modeling of Perpendicular STT-MTJs With Double Reference Layers," *IEEE Trans. on Nanotechnology*, vol. 18, pp. 1063–1070, 2019.
- [22] A. Jahanjoo, V. Bauer, S. Khooyooz, M. Haghi, and N. Taherinejad, "Real-Time and Energy-Efficient Attention Monitoring Using Single-Channel EEG for Everyday Life," *IEEE Access*, vol. 14, pp. 12 396–12 406, 2026.
- [23] P. Siirtola and J. Rönning, "Comparison of regression and classification models for user-independent and personal stress detection," *Sensors*, vol. 20, no. 16, 2020. [Online]. Available: <https://www.mdpi.com/1424-8220/20/16/4402>
- [24] A. Jahanjoo, N. TaheriNejad, and A. Aminifar, "High-Accuracy Stress Detection Using Wrist-Worn PPG Sensors," in *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2024, pp. 1–5.
- [25] K. Yang, B. Tag, Y. Gu, C. Wang, T. Dingler, G. Wadley, and J. Goncalves, "Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, ser. ICMR '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 562–570. [Online]. Available: <https://doi.org/10.1145/3512527.3531385>
- [26] S. Khooyooz, A. Jahanjoo, A. Aminifar, and N. TaheriNejad, "A Novel Machine-Learning-Based Noise Detection Method for Photoplethysmography Signals," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1–5.
- [27] M. A. F. Pimentel, A. E. W. Johnson, P. H. Charlton, D. Birrenkott, P. J. Watkinson, L. Tarassenko, and D. A. Clifton, "Toward a Robust Estimation of Respiratory Rate From Pulse Oximeters," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1914–1923, 2017.
- [28] R. Carboni, S. Ambrogio, W. Chen, M. Siddik, J. Harms, A. Lyle, W. Kula, G. Sandhu, and D. Ielmini, "Modeling of breakdown-limited endurance in spin-transfer torque magnetic memory under pulsed cycling regime," *IEEE Transactions on Electron Devices*, vol. 65, no. 6, pp. 2470–2478, 2018.
- [29] I. A. Danilov, M. S. Gorbunov, and A. A. Antonov, "SET Tolerance of 65nm CMOS Majority Voters: A Comparative Study," *IEEE Transactions on Nuclear Science*, vol. 61, no. 4, pp. 1597–1602, 2014.
- [30] B. Mohammad, K. Lin, P. Bassett, and A. Aziz, "A 65nm level-1 cache for mobile applications," in *Proc. IEEE Int. Conf. Microelectronics (ICM)*, 2008, pp. 5–10.