Supplementary Information

Trustworthy Tree-based Machine Learning by MoS_2 Flash-based Analog CAM with Inherent Soft Boundaries

Bo Wen[1,†], Guoyun Gao[1,†], Zhicheng Xu[1,2], Ruibin Mao[1], Xiaojuan Qi[1], X. Sharon Hu[3], Xunzhao Yin[2], and Can Li[1,4,*]

[1]Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR, China

[2]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China

[3]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA

[4]Center for Advanced Semiconductor and Integrated Circuit, The University of Hong Kong, Hong Kong SAR, China

[*]E-mail: canl@hku.hk

[†]These authors contributed equally to this work.

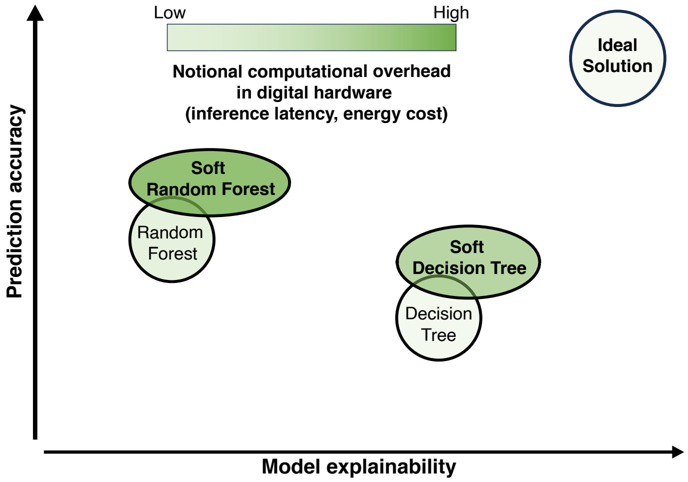

Fig. S1 | The prediction accuracy, model explainability, and computational overhead of several tree-based models. The shade of color represents the notional computational overhead in digital hardware. The soft tree-based models providing higher accuracy and explainability with probability-based decisions are inevitably accompanied by much heavier computational overhead than common tree-based models.

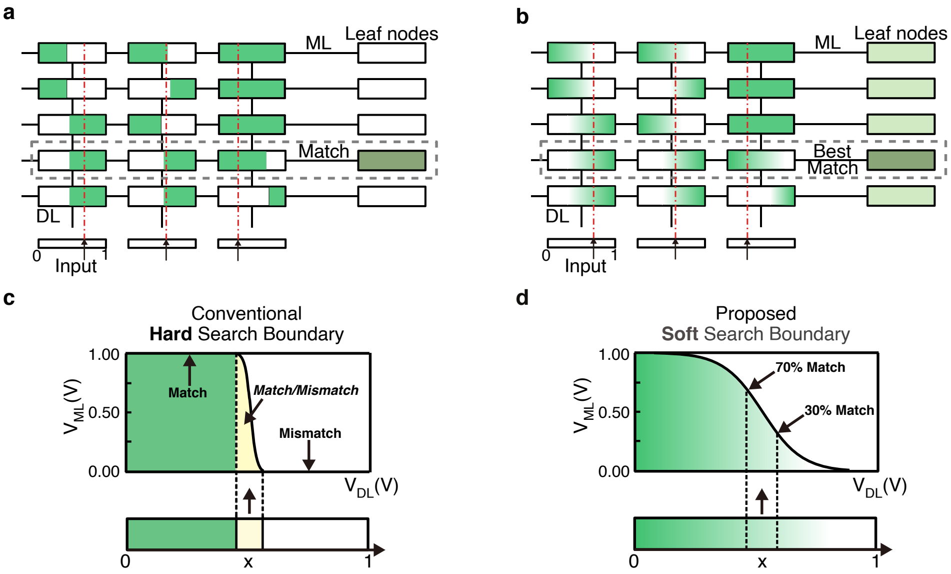

Fig. S2 | Comparison of proposed soft tree mapping with previous hard tree mapping. a, Analog CAM array performing the traditional hard DT, where each root-to-leaf path corresponds to a row of the array. The selected path is represented by the parallel searching result reflected on ML, as the only one match result. Tree-based models can be mapped and accelerated in analog CAM arrays. The data lines (DLs) accept input in form of voltages and output the results through match lines (MLs). **b**, Analog CAM array performing SDT, where the boundary are soft. The output is the best match leaf with the highest probability. **c**, The conventional search boundary in analog CAM is expected to be sharp enough for binary match or mismatch outputs. The boundary is not ideally orthogonal with inevitable intermediate state. **d**, The proposed search with soft boundary outputs continuous results with degrees of match rather than binary match or mismatch.

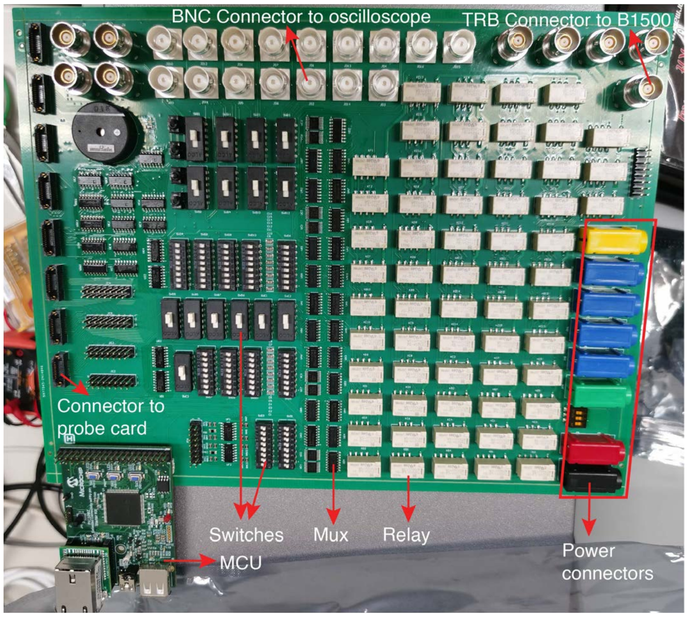

Fig. S3 | PCB-based test board for measurement platform. The multiplexers (MUXs) and relays on the test board are controlled by a Microchip PIC32 MCU connected with PC. The signals are collected by probe card and transferred through the board to B1500 for analysis. There are 16 connectors for oscilloscope to monitor the charge/discharge process of different analog CAM rows.

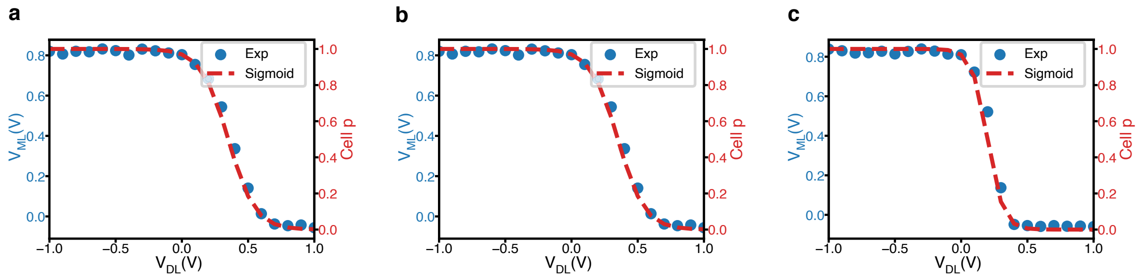

Fig. S4 | Sigmoid-like curves from ACAM cells. The sampled ML voltage well fitted by modified sigmoid functions for the other 3 cells (**a** | the 2nd cell. **b** | the 3rd cell. **c** | the 4th cell) on the row with their cell probability *p* which is similar to the cell showed in Fig.3b

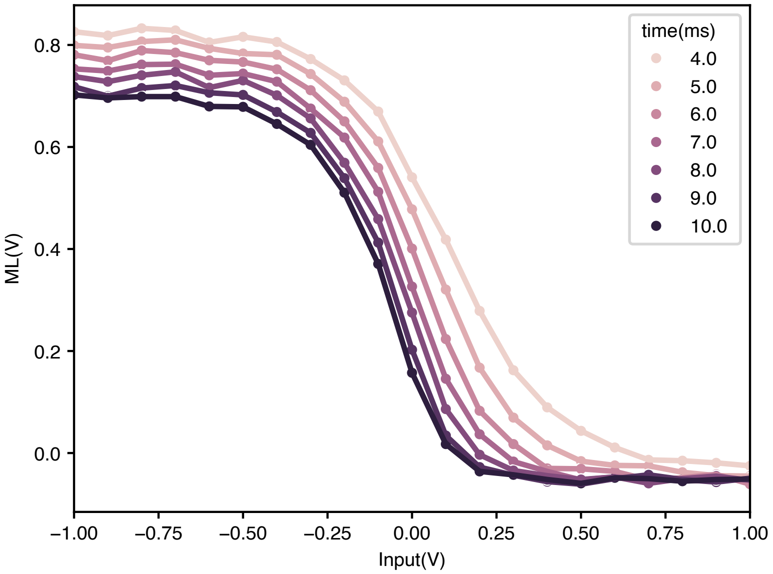

Fig. S5 | Sigmoid-like curves of ML voltage as output. The ML voltages are sampled at different time points from 4.0 ms to 10.0 ms. The curves get softer (flatter) with less time latency.

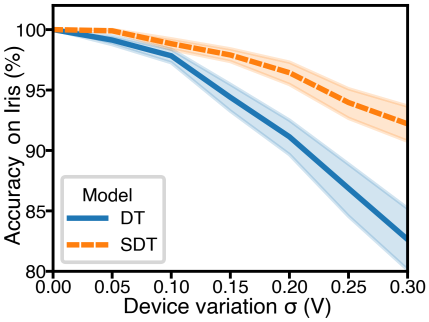

Fig. S6 | Simulation accuracy under Gaussian distributed device variation. SDT outperforms DT with increasing standard deviation σ. The trend is similar to results under uniform distributed device variation in Fig. 4a. When the σ is 0.3V, the accuracy of SDT is 92.2% ± 7.7% compared with 82.6% ± 12.8% of DT over 100 times of repetitions.

Model accuracy	DT			SDT		
Tree depth	Ideal	Root attack	Threshold variation	Ideal	Root attack	Threshold variation
6	74.15	54.79±0.49	35.23±8.85	75.66	67.08±0.18	74.8±0.88
8	81.81	66.43±0.25	42.44±7.38	84.31	78.08±0.23	83.22±0.4
10	86.62	72.94±0.32	41.99±6.77	88.12	83.93±0.24	87.53±0.18
12	87.74	73.51±0.35	46.36±6.03	90.1	87.43±0.18	89.52±0.19
14	88.17	73.93±0.19	42.47±3.1	90.77	88.23±0.25	90.04±0.21
16	88.11	73.94±0.23	45.27±6.56	90.79	88.4±0.13	90.27±0.23
18	88.02	74.05±0.22	40.37±4.38	91.15	88.98±0.13	90.75±0.17
20	88.26	73.92±0.23	42.99±4.56	91.26	89.62±0.12	90.69±0.18

Table S1 Model's accuracy with increasing tree depth under the root node adversarial attack / threshold variation. Every result is the average of 10 repetitions considering the randomness.

Dataset	#samples	#features	Tree depth	DT (%)	SDT (%)	RF (%)	SRF (%)
electricity	16714	10	6	76.59	79.08	78.55	78.35
bank-marketing	10578	7	6	78.22	80.06	80.62	80.15
default-of-credit-card-clients	13272	20	6	68.99	70.15	71.47	71.32
MiniBooNE	72998	50	6	88.97	91.37	90.23	90.36

Table S2 Tabular dataset information and ideal tree model accuracy. The 4 tabular datasets are from [1]

Process	Latency (ns)	Energy (nJ)	Reference
0.35 µm	48	3.62	[2]
0.18 µm	80	1.08	[3]
45 nm	3	0.07	[4]

Table S3 Latency and energy consumption of WTA for 3k inputs.

Model	Accelerator	Process	Power (mW)	Latency (ns/dec)	Energy (nJ/dec)
DT	AMD EPYC 7413	7 nm	180×10^3	0.6×10^3	0.11×10^6
SDT	AMD EPYC 7413	7 nm	180×10^3	0.3×10^6	54×10^6
SDT	NVIDIA RTX3090	8 nm	350×10^3	17.4×10^3	6.1×10^6
DT-based	ASIC IMC [5]	65 nm	7.1	2.7×10^3	19.4
DT	TCAM [6]	16 nm	247	3	0.74
DT-based	Memristive analog CAM [7]	65 nm	427	3	1.28
DT	This work	200 nm	675	10	8.78
SDT	This work w/o WTA	200 nm	675	10	8.78
SDT	This work w/ WTA	200 nm & 45 nm	681	13	8.85

Table S4 Latency and energy benchmark. For the analog CAM implementation, we consider a conservative latency estimation of 10 ns per array based on circuit simulation. The number can be further improved by adapting novel designs based on emerging FeFET technology[8]. We also consider the additional peripheral circuitry which is the WTA circuit[9]. That causes the latency for SDT slightly higher than DT.

References

1. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? Adv. neural information processing systems **35**, 507–520 (2022).

2. Fish, A., Milrud, V. & Yadid-Pecht, O. High-speed and high-precision current winner-take-all circuit. IEEE Transactions on Circuits Syst. II: Express Briefs **52**, 131–135 (2005).

3. Dlugosz, R., Rydlewski, A. & Talaśka, T. Novel, low power, nonlinear dilatation and erosion filters realized in the cmos technology. FACTA UNIVERSITATIS, Series: Electron. Energ. **28**, 237–249 (2015).

4. Liu, C.-K. et al. Cosime: Fefet based associative memory for in-memory cosine similarity search. In Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design, 1–9 (2022).

5. Kang, M., Gonugondla, S. K., Lim, S. & Shanbhag, N. R. A 19.4-nj/decision, 364-k decisions/s, in-memory random forest multi-class inference accelerator. IEEE J. Solid-State Circuits **53**, 2126–2135 (2018).

6. Rakka, M., Fouda, M. E., Kanj, R. & Kurdahi, F. Dt2cam: A decision tree to content addressable memory framework. IEEE Transactions on Emerg. Top. Comput. **11**, 805–810 (2023).

7. Pedretti, G. et al. Tree-based machine learning performed in-memory with memristive analog cam. Nat. communications **12**, 5806 (2021).

8. Yin, X. et al. Fecam: A universal compact digital and analog content addressable memory using ferroelectric. IEEE Trans. Electron Devices **67**, 2785–2792 (2020).

9. Liu, C.-K. et al. Cosime: Fefet based associative memory for in-memory cosine similarity search. In Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design, ICCAD ’22, 1–9 (Association for Computing Machinery, New York, NY, USA, 2022).