APPLIED SCIENCES AND ENGINEERING

Deep random forest with ferroelectric analog content addressable memory

Xunzhao Yin[1,2], Franz Müller[3], Ann Franchesca Laguna[4]*, Chao Li[1], Qingrong Huang[1], Zhiguo Shi[1,2], Maximilian Lederer[3], Nellie Laleni[3], Shan Deng[5], Zijian Zhao[5], Mohsen Imani[6], Yiyu Shi[5], Michael Niemier[5], Xiaobo Sharon Hu[5], Cheng Zhuo[1,2]*, Thomas Kämpfe[3]*, Kai Ni[5]*

Deep random forest (DRF), which combines deep learning and random forest, exhibits comparable accuracy, interpretability, low memory and computational overhead to deep neural networks (DNNs) in edge intelligence tasks. However, efficient DRF accelerator is lagging behind its DNN counterparts. The key to DRF acceleration lies in realizing the branch-split operation at decision nodes. In this work, we propose implementing DRF through associative searches realized with ferroelectric analog content addressable memory (ACAM). Utilizing only two ferroelectric field effect transistors (FeFETs), the ultra-compact ACAM cell performs energy-efficient branch-split operations by storing decision boundaries as analog polarization states in FeFETs. The DRF accelerator architecture and its model mapping to ACAM arrays are presented. The functionality, characteristics, and scalability of the FeFET ACAM DRF and its robustness against FeFET device non-idealities are validated in experiments and simulations. Evaluations show that the FeFET ACAM DRF accelerator achieves ~10^6×/10× and ~10^6×/2.5× improvements in energy and latency, respectively, compared to other DRF hardware implementations on state-of-the-art CPU/ReRAM.

INTRODUCTION

Edge intelligence in the era of Internet of Things (IoT) requires that raw data are analyzed locally instead of being transmitted back to the cloud for processing (*1*–*3*). Such edge intelligence can best be achieved by deploying an artificial intelligence (AI) hardware engine designed for IoT devices. Deep neural networks (DNNs) are highly effective in processing visual and speech data for various applications with high accuracy. However, DNN models face several fundamental challenges and are not readily deployable in the IoT. First, modern DNN models require large memories to store learned weights (commonly >1 GB) (*4*), well beyond the capacity of an embedded, on-chip memory in edge devices. External memories are therefore needed to store the entire DNN model. The requisite data transfers between on/off-chip memory lead to notable energy and latency overheads, which in turn limit the network complexities that may be deployed in edge devices. Second, to achieve high accuracy, DNNs require a huge amount of labeled training data. Data collection and preparation is expensive and time consuming for many tasks—especially for edge devices, considering their diverse functionalities and applications (*5*–*7*). Third, the "black box" nature and large parameter space of a DNN makes it challenging to analyze and understand how DNNs make their decisions. In certain domains, such as medicine, health care, and finance, the interpretability of a model is critical in establishing trust and developing solutions to other related problems (*8*–*11*). In light of these challenges, deep random forests (DRFs), a recently proposed interpretable and memory-efficient AI model (*12*), are considered to be an excellent alternative to DNNs in realizing lightweight AI engines for edge intelligence.

At a high level, DRF incorporates the core features of deep learning models, i.e., layer-by-layer processing, in-model feature transformation, and sufficient model complexity (*12*), as shown in Fig. 1A. DRF follows a cascaded structure where each layer in a DRF receives feature information extracted from the preceding level. Each layer is an ensemble of random forests (i.e., an ensemble of weak decision tree–based classifiers). Each forest models the class distribution of the datasets through either majority voting or averaging the predictions of decision trees in the same random forest. Those outputs from the forests in the same layer are concatenated together and forwarded to the next layer for further processing (*12*). Equipped with these deep model features, DRF achieves comparable or better accuracy with DNNs in processing low-resource dataset (*12*). In addition, by inheriting the interpretability and low energy and memory requirements of the random forest (*13*), DRF represents a competitive solution for edge intelligence to handle information processing tasks with requirements that DNNs might struggle to satisfy (e.g., limited resources or interpretability). Unlike DNNs, hardware acceleration of DRF has not been well explored. Our work addresses this gap by introducing an energy-efficient and high-performance hardware for accelerating DRF.

The key challenge in accelerating DRF is to implement the decision trees, the core component of DRF, as shown in Fig. 1B. It performs comparisons at each nonleaf node, and depending on the comparison results, the node is split into different branches. It has been proposed that analog content addressable memory (ACAM) can be used to perform the branch-split operation in a decision tree (*14*–*16*), which opens up the possibility of accelerating DRF with ACAM arrays. As a type of associative memories, content addressable memories (CAMs) have gained popularity in data-centric computing due to their massively parallel pattern-matching capability (*17*, *18*). They can identify the stored entries matching the search query in parallel in the exact or approximate matching mode. In the exact matching mode, only the items that exactly match the input query are identified (*18*), while in the approximate matching mode, the Hamming distance (HD) between the query and stored entries is returned by sensing the match-line (ML) current. The approximate matching function has been applied to accelerate various machine

[1]Zhejiang University, Hangzhou, Zhejiang, China. [2]Key Laboratory of CS&AUS of Zhejiang Province, Hangzhou, China. [3]Fraunhofer IPMS, Dresden, Germany. [4]De La Salle University, Manila, Philippines. [5]University of Notre Dame, Notre Dame, IN 46614, USA. [6]University of California, Irvine, CA 92697, USA.
*Corresponding author. Email: ann.laguna@dlsu.edu.ph (A.F.L.); czhuo@zju.edu.cn (C.Z.); thomas.kaempfe@ipms.fraunhofer.de (T.K.); kni@nd.edu (K.N.)

learning applications (*19*, *20*). All the developments above have only considered digital CAMs, where binary information is stored and searched. However, it is also possible to leverage the analog states of nonvolatile memories for multi-bit or ACAMs (*21*–*23*). Multi-bit information can be stored in the CAM, and an analog or multi-bit query can be searched across the CAM array for pattern matching, thus greatly improving the information density and expanding the CAM functionality (*21*–*24*). Here, we demonstrate ferroelectric ACAMs and leverage their unique properties to accelerate DRF.

In an ACAM cell, a matching range, defined by the upper and lower bounds of the search line (SL) voltage, can be dynamically adjusted by configuring the memory device states (*21*). We observe that by fixing the upper/lower bound of the matching range to the maximum/minimum voltage allowed on the SL and leaving the corresponding lower/upper bound adjustable, the respective greater-than (i.e., >)/less-than (i.e., <) branch-split operations in a decision tree can be efficiently implemented in an ACAM cell through a simple search operation, as shown in Fig. 1C. An ACAM word, composed of a row of CAM cells, can be used to implement a branch from the root node to a leaf node in a decision tree, while an ACAM array represents an entire decision tree. In this way, the decision space partitioned by the decision tree can be mapped into the

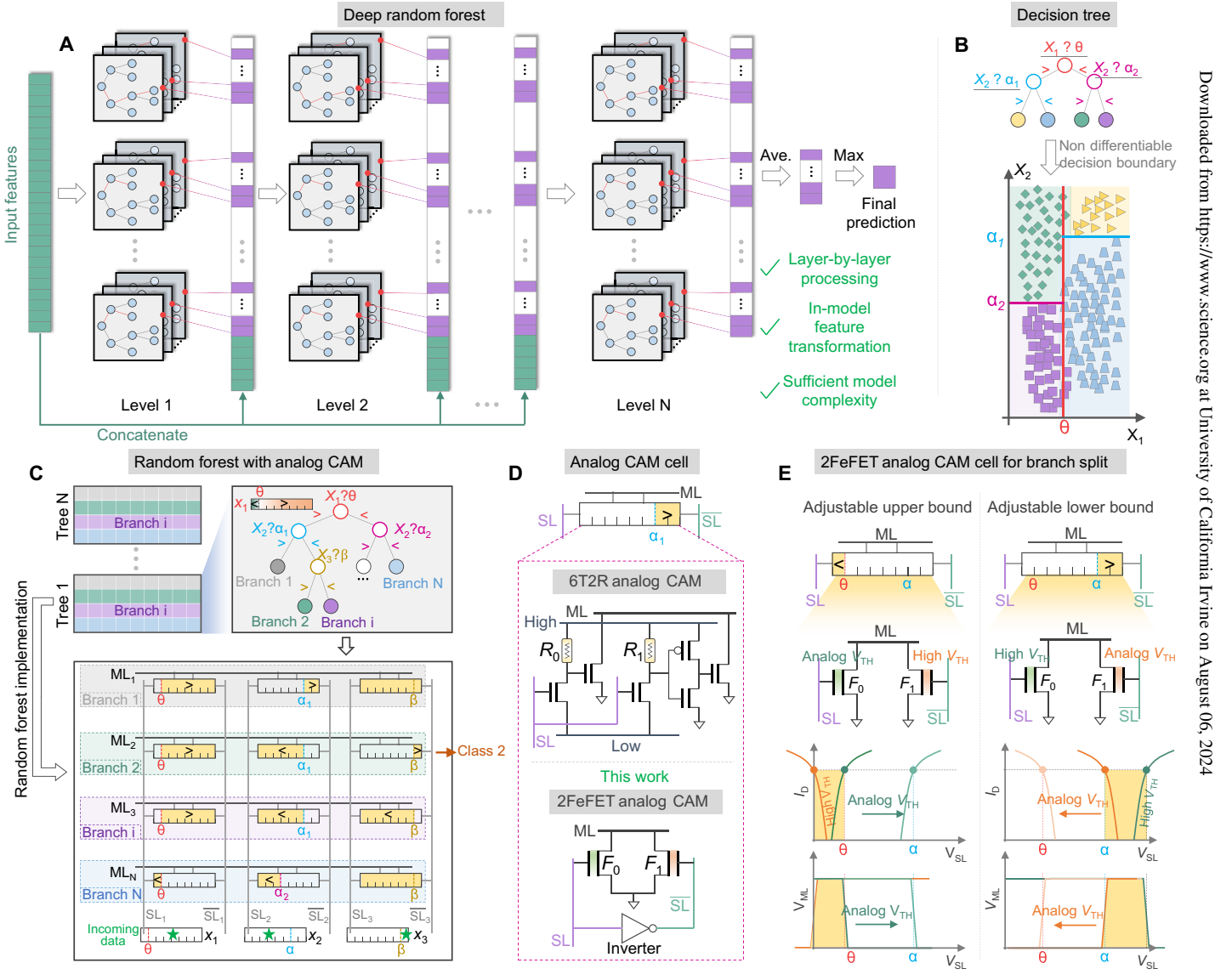

Fig. 1. Overview of implementing DRF with ferroelectric ACAM. (**A**) DRF is a deep model built by cascading random forests, forming a layer-by-layer structure. The output of each layer concatenates a portion of the input features, allowing in-model feature transformation. The resulting DRF model can achieve good performance. (**B**) Each decision tree in a random forest forms a nondifferentiable decision boundary by making a branch split at each nonleaf node based on the input features. (**C**) The random forest can be mapped onto an ACAM array. An ACAM cell with adjustable matching bounds (i.e., upper or lower matching bound) can efficiently realize the branch-split operation in a decision tree; as such, an ACAM word can realize a branch from the root node to the leaf node in a decision tree. (**D**) Existing demonstrated ACAM cells based on the multi-bit embedded nonvolatile memories. Compared with its 6T2R ReRAM ACAM counterpart, the 2FeFET-based ACAM is compact and universal by simultaneously serving as a digital and analog CAM. (**E**) Working principle of 2FeFET ACAM cell with adjustable upper/lower matching bound to realize the branch split in a decision tree.

matching space of an ACAM array. As a result, the inference operation of a decision tree can be realized through a simple parallel search operation in an ACAM array. The identified matched entries indicate the prediction results (i.e., the matching branches). By cascading multiple ACAM arrays together, the DRF can be realized. The effects of the limited precision of ACAM cells in defining the decision boundaries and the device-to-device variation of ACAMs are explored in the system benchmarking section.

Developing ACAM arrays for DRF requires that the ACAMs be compact, fast, and energy efficient. In our previous work (*21*), we have proposed a universal ferroelectric CAM design through SPICE simulations, in which a CAM cell composed of two ferroelectric field effect transistors (FeFETs) can simultaneously serve as a digital and analog CAM cell. Notably, the 2FeFET CAM is the most compact cell to date, compared with SRAM-based CAM cells typically composed of 16 transistors, spin-transfer-torque magnetic random access memory (STT-MRAM)–based CAM cells built using 10 to 15 transistors and 2 to 4 magnetic tunnel junctions (MTJs), and a resistive memory [i.e., resistive random access memory (ReRAM) and phase change memory (PCM)]–based CAM cell constructed with 2 transistors and 2 resistive memory devices (*20*). This compactness originates from its intrinsic three-terminal transistor structure and its capability in enabling single-FeFET cell memory array (*25*, *26*) when appropriate bias inhibition schemes are applied (*27*). Additionally, CAM based on FeFET is especially energy efficient. Unlike a volatile SRAM CAM, which consumes a substantial leakage power, FeFET CAM is nonvolatile, thus avoiding the energy consumption due to leakage current. Moreover, unlike other non-volatile memories (NVMs) where switching is typically driven by a large conduction current, ferroelectric switching can be induced with an applied electric field without consuming conduction current, thus exhibiting superior energy efficiency. Write energy down to 1 fJ/bit is achievable in a single FeFET (*2*, *28*). Finally, ferroelectric CAM exhibits superior performance owing to its intrinsic transistor structure and a large I_{ON}/I_{OFF} ratio (e.g., $\sim 10^4$), greatly outperforming the two-terminal resistive memories, which typically show an I_{ON}/I_{OFF} ratio of ~100. These characteristics enables 2FeFET CAM to simultaneously serve as both a digital and an analog CAM, creating a versatile hardware platform for various applications.

Here, we demonstrate the 2FeFET-based ACAM for the implementation of a DRF. There have been reports of using other NVM devices to implement an ACAM cell, such as the first proposed ReRAM ACAM (*22*) (Fig. 1D). However, due to its limited I_{ON}/I_{OFF} ratio, additional transistors are added into the digital CAM cell core (e.g., the 2T2R CAM cell) to support the analog/multi-bit search functionality, making it a 6T2R structure (*22*), larger than the 2FeFET ACAM design. The operating principles of the proposed 2FeFET ACAM cell for implementing the branch-split operation in a decision tree are illustrated in Fig. 1E. To implement a less-than branch (Fig. 1E, left), the FeFET F_1, connected with $\overline{SL}$, is set to the high-V_{TH} state such that it remains in the cutoff state over the entire SL search range, thus forming a fixed lower bound. Adjusting the V_{TH} state of the FeFET F_0 associated with the SL tunes the upper bound of the matching range. When the SL search voltage V_{SL} falls within the yellow region (where both the FeFETs turn off and the ML discharges slowly), the ML voltage, V_{ML}, remains high throughout the sensing phase of a voltage sense amplifier (SA). When the V_{TH} of F_0 increases, the resulting upper bound of the matching range also increases. As a result, the less-than branch with different thresholds can be mapped to the 2FeFET ACAM cell with an adjustable upper bound. By symmetry, the greater-than branch can also be achieved by setting F_0 in high-V_{TH} state, forming a fixed upper bound and adjusting the V_{TH} of F_1 to set the lower bound of the matching range. For the cases where not all the branches are of the same length, such as branch 1 and branch 2 in Fig. 1C, or of the same set of features for branch split, the "don't care" functionality of ACAM is leveraged. When a branch-split operation occurs over an input feature that is not included in the other branches, the ACAM cells mapping the missing features in those branches are set to the "don't care" state so that they contribute negligible leakage current through the ML, without affecting the V_{ML}. The "don't care" functionality can be realized by simply setting both FeFETs of the ACAM cell to the high-V_{TH} state.

In the following sections, we first describe the experimental demonstration of the 2FeFET ACAM cell and verify the branch-split operation for decision trees. We also demonstrate the capability of an ACAM word in realizing a branch from the root node to a leaf node in a decision tree through a simple search operation. This capability is used to realize a DRF, which exhibits good performance and superior energy efficiency. In addition, we present the evaluation of the impact of FeFET nonidealities, such as variation and limited precision, on the DRF performance to demonstrate the robustness of FeFET ACAM DRF. Compared with existing works, the major contributions of our work are as follows: (i) proposing a DRF accelerator leveraging the ferroelectric ACAM arrays for edge intelligence; (ii) first experimental demonstration of a ultra-compact, energy-efficient, and universal 2FeFET digital and analog CAM cell; (iii) first experimental demonstration of mapping a decision tree to an ACAM array, and first experimental demonstration of a DRF containing multiple layers; (iv) evaluating the impact of limited precision of the multi-bit matching ranges stored in FeFETs on the accuracy of DRF, and proposing a precision extension method using low-precision devices; (v) validating the robustness of FeFET ACAM–based DRF against device-to-device variation.

RESULTS

2FeFET analog CAM demonstration

In this section, we first discuss the experimental validation of the ACAM cell operation. We have constructed the proposed ACAM cell with the GlobalFoundries 28-nm high-κ metal gate (HKMG) FeFET technology (shown in Fig. 2, A and B). The device features an 8-nm-thick Si-doped HfO_2 ferroelectric thin film as the gate dielectric, capped with a TiN and polysilicon layer. A thin SiO_2 interlayer (~1 nm) is also present between the ferroelectric and the silicon substrate. Figure 2B shows the schematic cross-section of the device. Detailed process information can be found in (*29*). The local crystallographic phase has been characterized in the ferroelectric HfO_2 films by transmission–electron back-scattering diffraction (EBSD) (*30*), as shown in Fig. 2C. Dendritic grains consisting of the ferroelectric orthorhombic phase are observed, and only a small portion of the film grains are in the monoclinic dielectric phase, suggesting a good control over the ferroelectric phase through the high-temperature stressed annealing. This is consistent with our previous work in the analysis of the ferroelectric material stack on silicon, specifically focusing on the orientation distribution within the film (*31*), which also shows that the presence of the tetragonal phase is suppressed and a predominantly orthorhombic ferroelectric phase is achieved through careful optimization of the fabrication process and doping

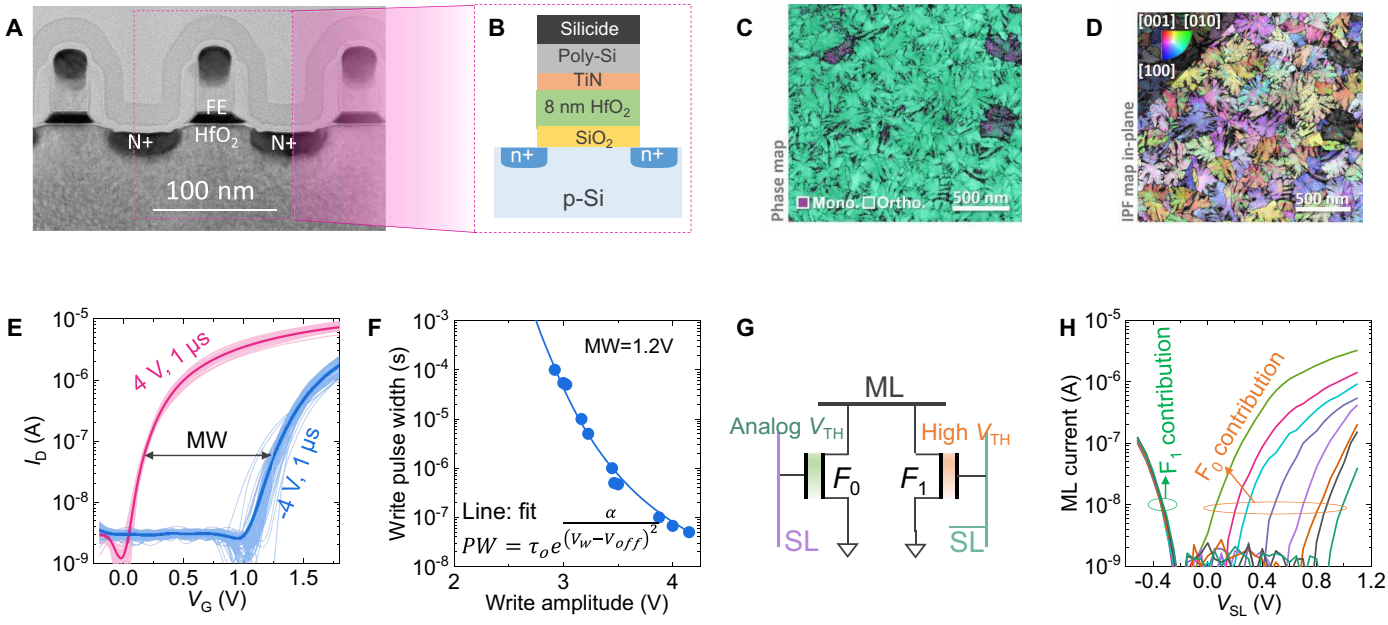

Fig. 2. Experimental demonstration of a ferroelectric ACAM cell. (**A**) Cross-sectional transmission electron microscopy (TEM) image of the FeFET device and (**B**) its schematic cross-section. It features an 8-nm-thick doped HfO_2 ferroelectric film. (**C**) The phase analysis through transmission-EBSD confirms that the poly-crystalline HfO_2 film consists mostly of the orthorhombic ferroelectric phase. Inverse pole figure maps (**D**) reveal intra-granular misorientation, especially in the dendrites. (**E**) Experimentally measured I_D-V_G characteristics for low-V_{TH} and high-V_{TH} states after ±4-V, 1-µs write pulses. Sixty different devices are measured, suggesting excellent device variation control in the FeFET. (**F**) Representative switching dynamics in the FeFET. To obtain a given memory window (e.g. 1.2 V in this case), the required switching time as a function of applied pulse amplitude can be well fitted with the nucleation limited switching model. (**G**) CAM cell configuration used in the experimental validation, where F_1 is set to be highest V_{TH} state and F_0 is adjusted. (**H**) Measured ML current as a function of the SL voltage, V_{SL}. Since F_1 is fixed to be highest V_{TH}, it contributes negligible current. When V_{TH} of F_0 is varied, the threshold of the matching range is shifted, thus demonstrating successful single-cell operation.

parameters. From the in-plane inverse pole figure map (Fig. 2D), a large variety of crystallographic orientations can be deduced. As a consequence, the polarization axis in each grain will be located at slightly different angles. Moreover, as gradients can be observed inside these grains and especially the dendrites, high degrees of intra-granular misorientation are expected (*32*). Consequently, these dendrites are likely to experience slightly different electric fields and are therefore reducing the effective grain size of the film. Note that a dendrite grain could contain different domains, and a broad distribution of polarization orientations, as present in this film, allows for analog-like multi-state operation in the ferroelectric HfO_2 layer.

The FeFET I_D-V_G characteristics for the low-V_{TH} and high-V_{TH} states after ±4-V, 1-µs write pulses are shown in Fig. 2E. Experimental setup for the cell and array measurement is shown in fig. S1. Device variation is characterized by measuring 60 different devices. The results show a large memory window of ~1.2 V and a large sensing margin (i.e., I_{ON}/I_{OFF}) separating the two V_{TH} states even when considering the device variation. The switching dynamics of the tested FeFET are shown in Fig. 2F, where the required pulse width to obtain a memory window of 1.2 V as a function of write pulse amplitude is presented. The required switching time can be well described by the expression derived from domain nucleation theory (*33*, *34*).

$$PW = \tau_o e^{\frac{\alpha}{(V_w - V_{off})^2}}$$

where α is a fitting parameter related with the polarization switching barrier, τ_o is the switching time at an infinitely large applied pulse amplitude, and V_{off} is the offset voltage, an indication of the local domain environment. With the increase of write pulse amplitude, FeFET switching speed can be further reduced to below 10 ns (*35*), suggesting the great promise for high-speed and energy-efficient ferroelectric memory.

Leveraging the partial polarization switching in the multi-domain FeFET, multiple V_{TH} states have been demonstrated and used for multilevel cell memories and synaptic weight cells for the acceleration of matrix-vector multiplication (*36*–*38*). Here, we harness the intermediate V_{TH} states to realize the branch-split operation with adjustable thresholds for the nonleaf nodes in a decision tree for DRF. Figure S2 shows experimentally measured I_D-V_G characteristics for four V_{TH} levels in a FeFET, which are set by applying different pulse amplitudes (*23*). The extracted V_{TH} distribution for four and eight levels are shown in fig. S2, C and D, respectively. With negligible overlaps between the neighboring levels, it is feasible to store multiple states into a FeFET, thus enabling the ACAM application proposed in this work. As shown in Fig. 2G, to verify the single ferroelectric ACAM cell operation, the FeFET associated with $\overline{SL}$ (F_1) is set to the high-V_{TH} state (V_{TH} = 1.1 V) and the FeFET associated with SL (F_0) is configured to different V_{TH} states. The ML current is then measured with a sweeping SL voltage, V_{SL}. As a result, the matching range of V_{SL} where the ML current is low can be identified. Such V_{TH} configurations define a matching range with varying upper bounds over the V_{SL}, thus implementing a less-than branch-split operation with varying decision boundaries. Because of the symmetry of the ACAM cell, the greater-than branch split is realized by simply swapping the V_{TH} settings of the two FeFETs. Figure 2H shows the measurement results corresponding to Fig. 2G. With the high-V_{TH} state of F_1, this

FeFET is cut off in the entire voltage range (i.e., 0 to 1 V) and is only turned on at negative V_{SL}. By setting V_{TH} of F_0 to eight different states, the upper bounds for the matching range are defined accordingly. As such, this verifies the successful operation of the ferroelectric ACAM cell.

To exploit a ferroelectric ACAM word for the mapping of an entire branch from the root node to a leaf node of a decision tree, we further validate the capability of an ACAM word to define a matching subspace in the high-dimensional feature space spanned by the V_{SL} inputs of all the ACAM cells. Figure 3A illustrates the experimental validation of the ferroelectric ACAM word. Figure 3B shows the compact layout for an ACAM word. Without loss of generality and for better illustration, an ACAM word consisting of two ACAM cells is demonstrated, which can define a matching subspace in the whole feature space spanned by V_{SL1} and V_{SL2}. For the experimental demonstration, similar to the single-cell case, the F_1 transistor in both cells is set to the high-V_{TH} state, while the V_{TH} state of F_0 is varied among four different levels from 0 to 1.1 V. In the ACAM array, each cell is independent from each other. As such, the V_{TH} of F_0 defines a V_{SL} plane, below which the cell contributes negligible current, indicating a match. When multiple cells are connected in parallel on the same ML, each cell defines one such V_{SL} plane, and the intersection of the space bounded by those planes defines the matching subspace of the ACAM word, namely, the search input space that satisfies all the split conditions along a branch of a decision tree. Figure 3C illustrates the ML current as a function of V_{SL1} and V_{SL2} when the

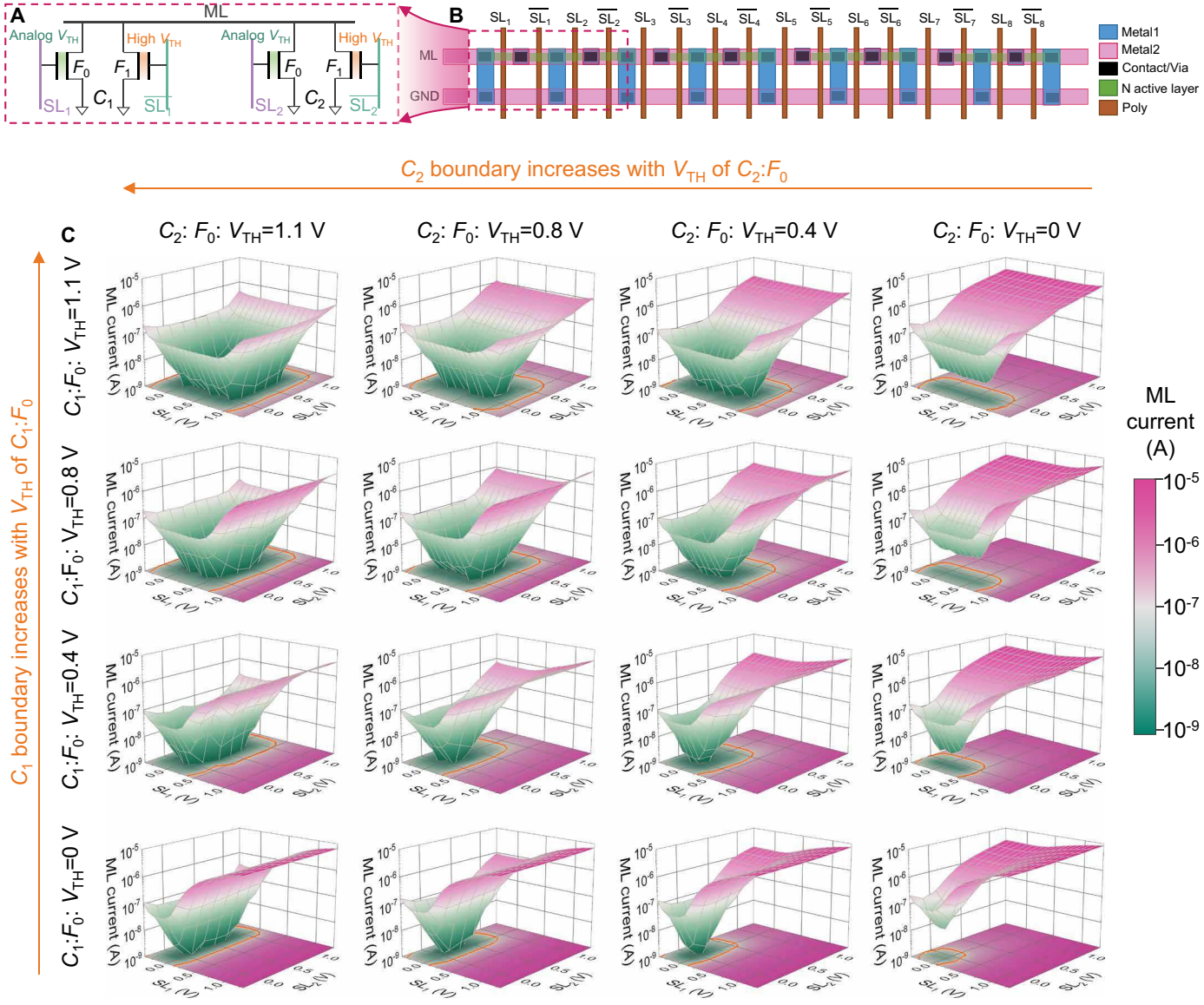

Fig. 3. Experimental demonstration of ferroelectric ACAM array. (**A**) Configuration of FeFETs in an ACAM word with two columns. F1 transistors in both cells are set to the high-V_{TH} state, and F_0 transistors in both cells are configured to different V_{TH} states, which set the threshold for the branch-split operation. (**B**) Compact layout of a 2FeFET ACAM word with a word length of 8. (**C**) The experimental results show that the low ML current region (i.e., matched condition) can be configured in different locations in the V_{SL} space. Orange lines in each figure correspond to a match line current of 10^{-7} A. It successfully demonstrates the capability of ferroelectric ACAM word in configuring the matching subspace in the overall V_{SL} space.

V_{TH} of F_0 in both cell 1 and cell 2 is set to one of the four different levels (a total of 4 × 4 configurations). The three-dimensional (3D) colormap surface of the ML current and its projection on the V_{SL1} and V_{SL2} plane are presented. It indicates that the low current region on each dimension (e.g., $\lesssim 10^{-7}$A in this work) follows the V_{TH} states of the F_0 transistor in the corresponding cell. This successfully demonstrates the independence among the ACAM cells. Thus, the configured cell threshold sets the boundary of the matching subspace on the dimension of the corresponding cell.

An ACAM array with a larger word length of 16 has also been tested. As the matching subspace of the word lies in the 16D space and cannot be visualized, for ease of illustration, we consider a configuration where 15 cells are grouped together by storing the same state and are searched with the same information. The F_1 transistors in all ACAM cells are in the high-V_{TH} state, enabling all cells to perform the lower-than branch-split operation. The F_0 transistors in the grouped 15 cells are set to the same intermediate state. The remaining single cell is adjusted among the four different V_{TH} states. After configuring the cells, the V_{SL} of the single cell and that of the grouped cells are swept from −0.3 V to 1.2 V in steps of 0.1 V. Figures S3 to S6 show the measured ML current when the F_0 transistors of the grouped 15 cells are set to V_{TH} = 1.1, 0.8, 0.4, and 0 V, respectively. It shows that on the dimension of each V_{SL}, the low ML current matching range closely follows the V_{TH} of the corresponding cell. This indicates that the boundary of the matching subspace on one V_{SL} dimension in the high-dimensional space is set by the decision boundary of that particular ACAM cell. This verifies the basic operation principles of the proposed ACAM array in realizing the branch-split operation of a decision tree in a DRF.

Experimental verification of an ACAM array is also conducted. Figure 4A shows a decision tree composed of two layers and four leaf nodes. It can be mapped to an ACAM array with two columns and four rows, as shown in Fig. 4B. Programming of the FeFET V_{TH} states in the array is shown in Fig. 4C, and the programmed V_{TH} states are close to target values. On the basis of the mapping, the theoretical and the resulted experimental matching space partitioned by the two SLs are shown in Fig. 4, D and E, respectively. The experimental results resemble the theoretical ones, with wider boundaries between neighbor matching regions. The uncertainty at the boundary regions can be reduced by optimizing the subthreshold swing of FeFETs. In addition, an example DRF composed of three layers and one decision tree per layer is experimentally constructed and verified, as shown in fig. S7. The results show that the experimental measurements match the theoretical results, indicating the successful DRF operation with the ACAM array. These preliminary experimental demonstrations indicate that the proposed hardware-algorithm co-design solution is promising toward practical applications.

To use an ACAM array, voltage domain sensing is typically adopted for its simplicity, where the SA output voltage remains high when the search information matches the stored ACAM word; otherwise, the ML voltage discharges to ground. Such functionality has also been validated in SPICE simulations using a calibrated FeFET compact model (*39*), as shown in Fig. 4. Here, a single two-stage buffer circuit is adopted for voltage domain sensing, where the output is binary, as shown in fig. S8. The output is close to V_{DD} when a low current flows through the ML (i.e., match case) and at ground when a mismatch happens. Figure 4F shows the simulated ML current of a single cell configured to perform the less-than branch-split operation. The simulated ML current shows a similar trend as the experimental results shown in Fig. 2H. With this ML current dependence on V_{SL}, voltage domain sensing can be performed with the SA shown in fig. S8A. The simulated output transient waveforms at different search voltages are shown in fig. S14. For V_{SL} in the matching subspace, the ML current is low; thus, ML voltage remains high. Otherwise, the ML voltage discharges to ground at a fast rate. At a certain sense time (e.g., in this work, 10 ns is chosen), the SA output voltage varies as a function of V_{SL}, and multiple voltage thresholds for the branch-split operation can be defined depending on the stored V_{TH} in the cell, as shown in Fig. 4G. Therefore, whether the input query matches with the defined branch condition can be determined by the output of the SA.

The operations of the ACAM array are also simulated. Similar to the experiment shown in Fig. 3, the ML current of an ACAM array with two columns is simulated by sweeping V_{SL1} and V_{SL2} of the cells. By setting V_{TH} of F_0 in both cells in one of 4 × 4 configurations, different match subspaces can be realized in the space spanned by V_{SL1} and V_{SL2} (as shown in fig. S15, following the same behavior as the experiment shown in Fig. 3). Voltage domain sensing of the ACAM array is also implemented using the same setup as the single cell as shown in fig. S8B. The impact of the array size (i.e., rows and columns of the ACAM array) on the voltage sensing of the ACAM array has been simulated, as shown in Fig. 4 (H to K). A worst-case scenario is considered, where only one cell in the array is swept while all the other cells are searched with a V_{SL} close to the decision boundary, which makes it challenging to sense. The impact of the number of columns (i.e., the number of ACAM cells connected to the same match line) on the sensing of the ACAM array is studied. As the number of cells per word increases, the leakage current contributed by the cells searched close to the boundary becomes larger, resulting in an increased discharge rate of the match line. Therefore, as shown in fig. S16, when the column size increases from 1 to 32, the search time needs to be adjusted accordingly. Figure 4J shows the output voltage as a function of V_{SL1} for ACAM arrays with different number of columns sensed at the search times shown in Fig. 4H. It can be seen that the decision boundary can be maintained across various sizes of arrays. Since the array size is predetermined, the adjustment of sense time is straightforward. Figure 4 (I and K) shows that the impact of the number of rows, i.e., number of ACAM words or independent match lines, on the array sensing is negligible, as each ML sensing is independent. Therefore, the decision boundary can be maintained for scaled array sizes.

Application evaluation and benchmarking

Leveraging the validated FeFET ACAM array, the performance of DRF can be evaluated. The mapping of a DRF involves multiple ACAM arrays. As demonstrated in Fig. 1, DRF is a machine learning framework that follows a layer-by-layer structure using cascaded random forests. Each layer is composed of multiple random forests, which output a probability for each class. A random forest uses an ensemble of decision trees to determine the probability of each target class for a given test example. Each decision tree can be mapped to an ACAM array as shown in Fig. 5A. Each cell represents a nonleaf node that performs the branch-split operation over a specific feature. Each row of the ACAM implements a branch from the root node to a leaf node. Hence, the number of rows corresponds to the number of leaf nodes (i.e., number of branches). The number of columns in an ACAM array corresponds to the number of features. Multiple ACAM arrays can be cascaded horizontally as shown in Fig. 5A to hold all the features of a decision tree. As each cell in an ACAM

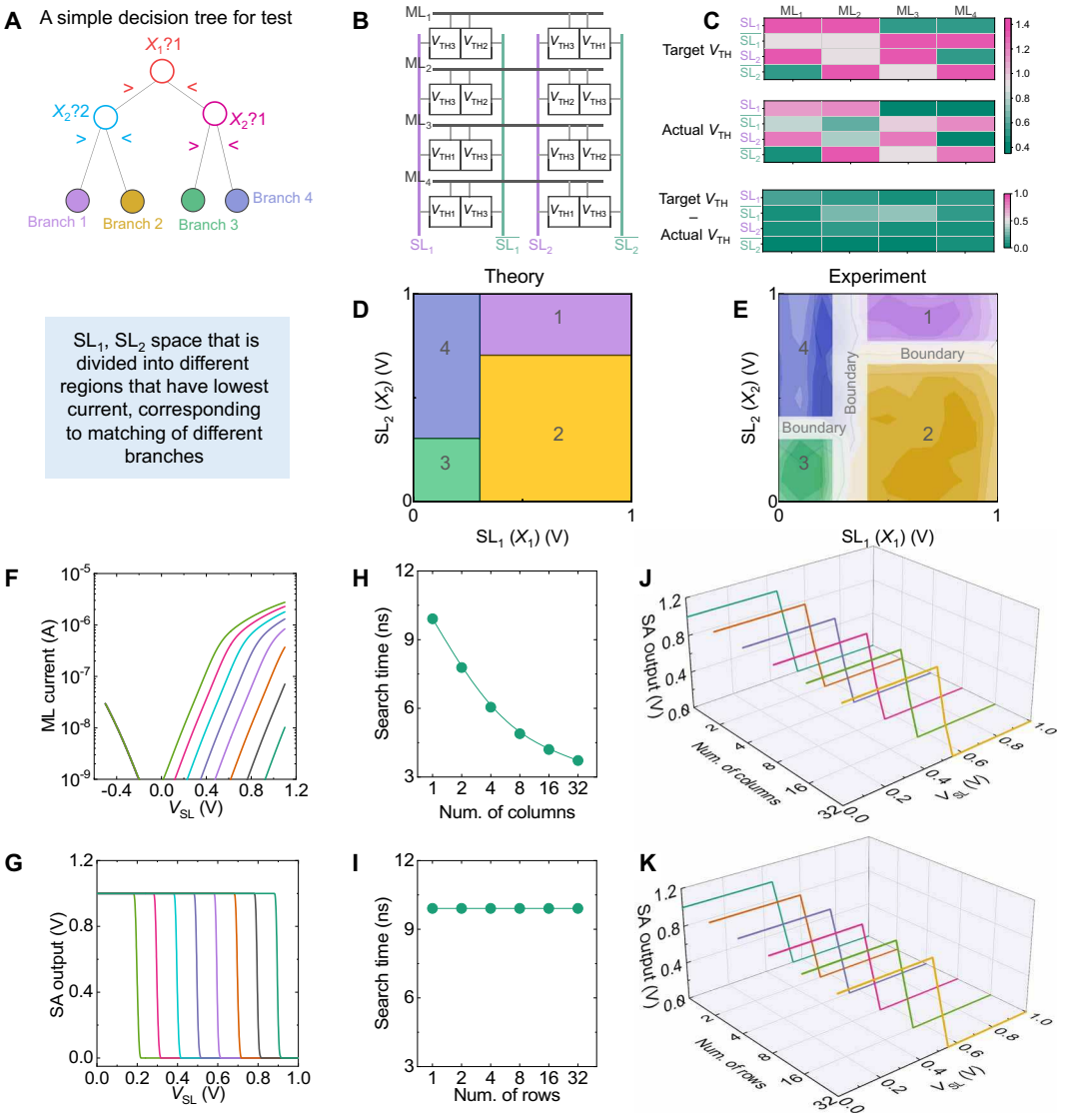

Fig. 4. Experimental and simulation validations of the proposed ACAM array. (**A**) Decision tree adopted for the experimental validation. (**B**) ACAM array implementing the decision tree. (**C**) Experimental V_{TH} programming of the ACAM array and the programmed states versus target states. (**D** and **E**) Theoretically and experimentally obtained matching space for the decision tree, with respect to SL_1 and SL_2, respectively. SL_1 and SL_2 space is divided into four regions corresponding to the four matching ranges stored in the CAM array rows in (B), thus mapping the four branches in (A). (**F**) Simulated ML current of a single ACAM cell with different V_{SL} when the ACAM cell is configured to perform the less-than branch-split operation. (**G**) Simulated transfer characteristics of the two-stage buffer SA over the search input voltage at the search time of 10 ns for the less-than branch-split operation. (**H** and **I**) The corresponding search time needs to be adjusted for different number of columns (H), while the search time is almost the same for different number of rows (I). (**J** and **K**) Transfer characteristics of the SA output over the input voltage as a function of the number of columns and rows in the ACAM array, respectively.

word is independent of each other, a large ACAM word can be decomposed into multiple small ACAM words such that searching for a large ACAM matching word is equivalent to searching for the matching words in all ACAM subarrays simultaneously. Each ACAM array corresponding to a decision tree votes for a given class, and using a vote counter, the random forest outputs a vote vector, which represents the number of votes for each class.

The vote vectors of the random forests are then concatenated and passed to the next layer of the DRF. The specific circuit details, including the decision tree implementation, the sensing circuitry, the digital-to-analog converter (DAC) for SLs, the random forest implementation, and the digital logic to postprocessing the match/mismatch, performing the majority voting for the forest and transmitting the intermediate results between the forests, are discussed in figs. S9 to S13, respectively. Here, the DRF simulation is implemented on an Intel Core i7-10750H 6-Core CPU and a Titan X GPU. The branch-split threshold is quantified using linear quantization. In our experiments, up to eight layers, eight random forests per layer, and 64 trees per forest are used.

DRFs have been used in a variety of applications such as facial age estimation (*40*), malware detection (*41*), and classification of hyperspectral images (*42*). Here, we use two representative datasets for benchmarking to evaluate the accuracy of the DRF model. One is an image dataset, MNIST (*43*), and the other is the time-series

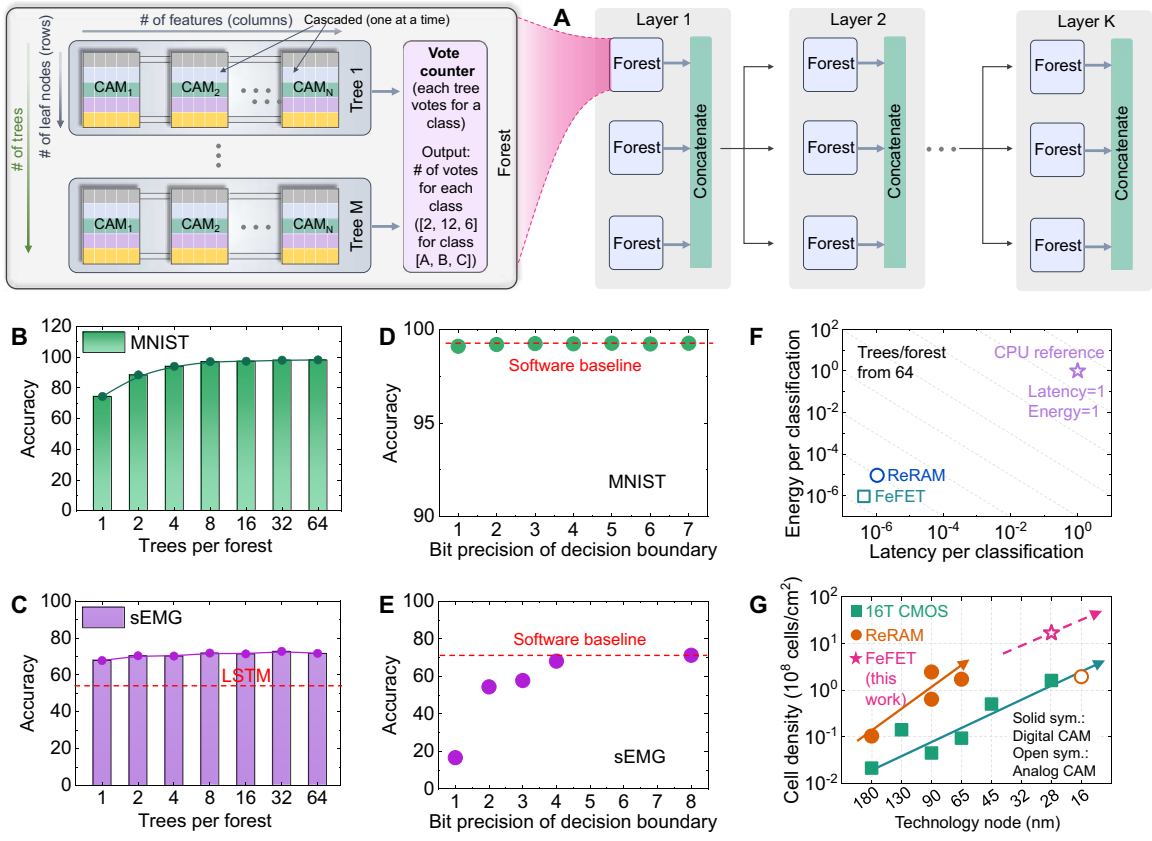

Fig. 5. Benchmarking of the DRF using ferroelectric ACAM arrays. (**A**) Mapping of the DRF onto ferroelectric ACAM arrays. Each tree of a forest is mapped to an ACAM array, where the number of rows corresponds to the number of leaf nodes (i.e., branches) and the number of columns corresponds to the total required features. (**B** and **C**) Inference accuracy for the MNIST and sEMG dataset with respect to the number of trees per forest, respectively. Excellent accuracy is obtained with the DRF, even when compared with the LSTM models. (**D** and **E**) Accuracy when mapped to the ACAM array considering the limited precision of the branch-split decision boundary. (**F**) Energy versus latency for a single classification using the DRF when mapped to the CPU, ReRAM- and FeFET-based ACAMs, respectively. FeFET-based ACAM shows superior performance. (**G**) ACAM cell density, including both the digital and analog cells. The 2FeFET-based ACAM achieves the highest density due to its compactness.

dataset, sEMG, used for hand movement recognition (*44*). The sEMG dataset consists of 1800 records, where each one belongs to one of six hand movements, i.e., spherical, tip, palmar, lateral, cylindrical, and hook. Each record captures approximately 6 s at 500 samples per second or approximately 3000 samples. For the sEMG data, we used a sliding window approach with a window size of 100 samples and a 95% overlap. This results in a 100 × 600 dimensional vector, which is used as an input to the DRF. Figure 5 (B and C) shows the inference accuracy for the MNIST and sEMG dataset as a function of the number of trees per forest in the DRF. We follow the training procedure in (*12*) while varying the number of trees. The DRF is trained at full precision, and the branch-split decision boundary is quantized after training to evaluate the impact of the boundary precision.

For both models, the accuracy saturates when more than eight trees per forest are used. An accuracy of 99.2% is achievable for the MNIST dataset, which is on par with a three-layer convolutional deep belief network (*45*). For sEMG, the accuracy of the DRF model is 72%, substantially outperforming an advanced long short-term memory (LSTM) machine learning model (*12*). These results demonstrate the competitive performance of DRF in performing different classification tasks.

As FeFET ACAM cell can currently hold three bits of V_{TH} states in this work (per fig. S2), the impact of precision on inference accuracy is evaluated. Figure 5 (D and E) shows the inference accuracy as a function of precision of the decision boundary for the MNIST and sEMG dataset, respectively. For MNIST, each grayscale pixel intensity is used as a feature, i.e., nonleaf branch-split node. Since relevant features are either black or white, the DRF performs well even at 1-bit precision. However, for the sEMG dataset, the accuracy starts to degrade when the decision boundary precision drops below 4 bits and accuracy is especially low at 1-bit precision. The FeFET ACAM with 3-bit precision demonstrated in this work suffers accuracy degradation but still performs better than LSTM for the sEMG dataset. Note that due to the core tree structure in a DRF, a higher precision branch-split operation can be realized using ACAM cells with lower precision at the cost of additional ACAM area and energy consumption. To implement a higher precision DRF, each tree node or some critical nodes (i.e., requiring a higher precision) can be split into multiple tree nodes (lower precision), as illustrated in fig. S17 as an example. Each feature must be split into its most significant bits (MSBs) and least significant bits (LSBs) and treated as two separate features and searched separately. This results in an increased number of branches, and hence the number of rows when mapping to ACAM

arrays. In future work, we will evaluate the trade-offs of extending the precision for FeFET ACAMs.

It is also important to evaluate the impact of device-to-device variation of FeFETs on the classification accuracy of the DRF. The variation in FeFET V_{TH} (per fig. S2) is directly translated into the variation in the decision boundary, which impacts the accuracy of the branch-split operations. As V_{TH} variation increases, overlap between neighboring decision boundaries is expected. The impact of such variations may vary by datasets. For MNIST, because the input is binary, the DRF is highly robust to variation as long as the decision boundary between the black and white pixels is well defined. For sEMG, the input values are not binary, but intermediate values, which increase the susceptibility of the system to FeFET variation. However, as suggested in fig. S18, when the SD of the decision boundary is less than 7% of the overall memory window, the accuracy remains unaffected. Considering that current FeFET V_{TH} SD is on average 4% of the overall memory window, DRF that leverages even current devices still yields negligible accuracy loss, demonstrating great robustness. As the FeFET technology continues to improve, variations will be further suppressed (*46*); thus, FeFETs will become an even more robust technology platform for DRF implementation.

To compare FeFET ACAM–based DRF with alternative DRF implementations, the ferroelectric ACAM array performance extracted from the simulations in Fig. 4 is used for system-level benchmarking. We assume an ACAM array of size 128 × 128 as the basic ACAM module and that multiple ACAM arrays are cascaded to complete all system-level tasks. Figure 5F shows energy versus latency for a single classification. The DRF implementation on an Intel Core i7-10750H CPU (14-nm node) @ 2.60 GHz with 16 GB of RAM is used as a reference (i.e., the latency and energy per classification is considered as 1), against which the system implementation using ACAM arrays based on ReRAM [16-nm node (*14*)] and FeFETs is benchmarked. The instantaneous CPU power during the inference of a DRF model is extracted by Power API of an Intel i7-10750H CPU. The energy per classification is calculated using the average power and the latency. The extracted FeFET ACAM array energy includes the ML precharge energy, the SA energy associated with the MLs, and the DACs driving the SLs, as shown in fig. S19. The search latency is determined, similar to fig. S14, as the time point when the corresponding matching boundary of the SA output transfer curve with V_{SL} aligns with the predefined boundary, i.e., the stored matching boundary. Since ReRAM ACAM array has only been proposed for a decision tree implementation and not for DRF (*14*), we take the reported ReRAM ACAM array characteristics, demonstrating outstanding performance gain against digital processors, and evaluate its performance in implementing the DRF. Because of their parallel nature and compact, in-memory computing characteristics, the ferroelectric ACAM array exhibits notable savings in energy and latency when compared with a CPU (e.g., up to 10^6× saving in energy and latency). FeFET-based ACAM arrays have lower energy consumption than their ReRAM counterpart due to the elimination of the DC flowing through the ReRAM ACAM cell. These results suggest great promise for the ferroelectric ACAM array when implementing the DRF. In addition, we also implemented a simple random forest model (i.e., no layer-by-layer structure) using the ferroelectric ACAM and evaluated its performance on some electroencephalogram (*47*) and positron emission tomography (*48*) dataset. Table S1 summarizes the metrics including cell size, energy, and latency per classification using our ferroelectric ACAM–based random forest, as well as other advanced machine learning model implementations. Again, superior energy efficiency and latency for a classification operation using the ferroelectric analog CAM array is demonstrated.

Figure 5G provides the evolution of CAM cell density as a function of technology nodes. Both the digital and analog CAM cells are included for completeness. As expected, with technology scaling, CAM cell density continues to improve. Because of its compactness, the ferroelectric ACAM cell (2FeFET) exhibits the highest density so far, greatly outperforming its ReRAM counterpart (6T2R). As a result, the compact ferroelectric ACAM array could well support the acceleration of the DRF model.

DISCUSSION

Here, we implemented the DRF with ferroelectric ACAM array by leveraging the parallelism and in-memory computing capability of the ACAM array. We demonstrated that DRF inference could be efficiently mapped as the associative search operations in ACAM arrays, as the ACAM cell can realize the key branch-split operation of a decision tree in memory by harnessing the analog polarization states within an FeFET. We validated the functionality of the 2FeFET ACAM cell and the capability of ACAM arrays in identifying the matching region in the high-dimensional search space. Each ACAM row corresponds to a specific branch from the root node to a leaf node in a decision tree. With the proposed ultra-compact ACAM cell, we show that the FeFET ACAM–based DRF accelerator exhibits order-of-magnitude improvement in footprint, and inference energy and latency. These results suggest that ferroelectric ACAMs provide a promising hardware platform to implement DRF as an alternative complement to DNNs for achieving edge intelligence with its interpretability, low latency, and superior energy efficiency.

MATERIALS AND METHODS

Device fabrication

Here, the fabricated FeFET features a poly-crystalline Si/TiN (2 nm)/doped HfO_2 (8 nm)/SiO_2 (1 nm)/p-Si gate stack. The devices were fabricated using the GlobalFoundries 28-nm node gate-first HKMG complementary metal-oxide semiconductor (CMOS) process on 300-mm silicon wafers. The ferroelectric gate stack process module starts with growth of a thin SiO_2-based interfacial layer, followed by the deposition of an 8-nm-thick Si-doped HfO_2. A TiN metal gate electrode was deposited using physical vapor deposition (PVD), on top of which the poly-Si gate electrode is deposited. The source and drain n+ regions were obtained by phosphorus ion implantation, which were then activated by a rapid thermal annealing (RTA) at approximately 1000°C. This step also results in the formation of the ferroelectric orthorhombic phase within the doped HfO_2. For all the devices electrically characterized, they all have the same gate length and width dimensions of 1 μm × 1 μm, respectively.

Electrical characterization

The FeFET device characterization was performed with a PXI-Express system from National Instruments, using a PXIe-1095 cassis, NI PXIe-8880 controller, NI PXIe-6570 pin parametric measurement unit (PPMU), and NI PXIe-4143 source measure unit (SMU). Before characterization, all FeFETs are preconditioned using the SMUs by cycling them 100 times with the pulses of +4.5 V, −5 V with a pulse length of 500 ns each. Readout of the memory state is done by a

stepwise increase of the gate voltage in 0.1 V-increments while applying 0.1 V to the drain terminal and measuring the current using the PPMU. Bulk and source terminals are tied to ground at all times. The read operation takes approximately 7 ms. The multilevel characterization of individual FeFETs is performed by putting them in a reference state with a gate voltage of −5 V or + 4.5 V for 500 ns for erase or program, respectively. After that, a single pulse of increasing amplitude is applied for 200 ns. The gate voltage amplitude stepping is set to 100 mV. After each pulse, a delay of 2 s is added to ensure sufficient time for charge detrapping after which a readout is performed. This scheme is repeated for the full switching range. The CAM measurements are performed in an AND-connected array. One CAM cell is constructed by measuring two FeFETs sharing the same connection at their drain terminal, the matchline. Source and bulk terminal are tied to ground at all times. The FeFETs are programmed to the target V_T's individually, applying a single fixed program pulse specific to the target V_T. Readout operation is performed similar to the single devices. The ML is kept at 0.1 V, while an stepped gate sweep is performed. Using individual PPMU channels, the readout is performed on both FeFETs of one CAM cell.

Transmission-EBSD characterization

For transmission-EBSD characterization, also known as transmission Kikuchi diffraction, a 10-nm Si-doped HfO_2 layer was deposited on a silicon wafer with a thin chemical oxide layer. This was carried out using atomic layer deposition with a cycling ratio of 16:1 (Hf:Si). After capping the layer with a 10-nm TiN top electrode, the film was crystallized via RTA at 1000°C. A dimpled sample was prepared and analyzed in a scanning electron microscope using a Bruker Optimus TKD detector. An acceleration voltage of 30 kV and a current of 3.2 nA were used.

SPICE simulation setup

The FeFET compact model (*39*) adopted in our simulations is calibrated by the fabricated FeFET, along with the predictive technology model (PTM) for CMOS transistors (*49*). The supply voltage is 1 V, and the temperature is set to 27°C. Wiring parasitics are extracted from DESTINY (*50*). For an ACAM array of size 128 × 128, 6.4- and 3.84-fF parasitic capacitance are extracted and associated with MLs and SLs, respectively.

Supplementary Materials

This PDF file includes:

Figs. S1 to S19
Table S1
References

REFERENCES AND NOTES

1. A. Keshavarzi, W. van den Hoek, Edge intelligence—On the challenging road to a trillion smart connected IoT devices. *IEEE Des. Test* **36**, 41–64 (2019).
2. A. Keshavarzi, K. Ni, W. van den Hoek, S. Datta, A. Raychowdhury, Ferroelectronics for edge intelligence. *IEEE Micro* **40**, 33–48 (2020).
3. Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, J. Zhang, Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proc. IEEE* **107**, 1738–1762 (2019).
4. X. Xu, Y. Ding, S. X. Hu, M. Niemier, J. Cong, Y. Hu, Y. Shi, Scaling for edge inference of deep neural networks. *Nat. Electron.* **1**, 216–222 (2018).
5. X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, X. Chen, Convergence of edge computing and deep learning: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **22**, 869–904 (2020).
6. D. Xu, T. Li, Y. Li, X. Su, S. Tarkoma, T. Jiang, J. Crowcroft, P. Hui, Edge intelligence: Architectures, challenges, and applications. arXiv:2003.12172 [cs.NI] (2020).
7. Z. Feng, S. George, J. Harkes, P. Pillai, R. Klatzky, M. Satyanarayanan, Edge-based discovery of training data for machine learning, in *2018 IEEE/ACM Symposium on Edge Computing (SEC)* (IEEE, 2018), pp. 145–158 .
8. F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning. arXiv:1702.08608 [stat.ML] (2017).
9. S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, P. Gurram, Interpretability of deep learning models: A survey of results, in *Proceedings of the 2017 IEEE Smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)* (IEEE, 2017), pp. 1–6.
10. A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.* **32**, 18069–18083 (2020).
11. A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).
12. Z.-H. Zhou, J. Feng, Deep forest. *Natl. Sci. Rev.* **6**, 74–86 (2019).
13. M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15**, 3133–3181 (2014).
14. G. Pedretti, C. E. Graves, S. Serebryakov, R. Mao, X. Sheng, M. Foltin, C. Li, J. P. Strachan, Tree-based machine learning performed in-memory with memristive analog cam. *Nat. Commun.* **12**, 5806 (2021).
15. G. Pedretti, S. Serebryakov, J. P. Strachan, C. E. Graves, A general tree-based machine learning accelerator with memristive analog CAM, in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2022), pp. 220–224.
16. G. Pedretti, J. Moon, P. Bruel, S. Serebryakov, R. M. Roth, L. Buonanno, A. Gajjar, T. Ziegler, C. Xu, M. Foltin, P. Faraboschi, J. Ignowski, C. E. Graves, X-time: An in-memory engine for accelerating machine learning on tabular data with CAMs. arXiv:2304.01285 [cs.LG] (2023).
17. K. Pagiamtzis, A. Sheikholeslami, Content-addressable memory (CAM) circuits and architectures: A tutorial and survey. *IEEE J. Solid-State Circuits* **41**, 712–727 (2006).
18. R. Karam, R. Puri, S. Ghosh, S. Bhunia, Emerging trends in design and applications of memory-based computing and content-addressable memories. *Proc. IEEE* **103**, 1311–1330 (2015).
19. M. Imani, A. Rahimi, D. Kong, T. Rosing, J. M. Rabaey, Exploring hyperdimensional associative memory, in *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)* (IEEE, 2017), pp. 445–456.
20. K. Ni, X. Yin, A. F. Laguna, S. Joshi, S. Dünkel, M. Trentzsch, J. Müller, S. Beyer, M. Niemier, X. S. Hu, S. Datta, Ferroelectric ternary content-addressable memory for one-shot learning. *Nat. Electron.* **2**, 521–529 (2019).
21. X. Yin, C. Li, Q. Huang, L. Zhang, M. Niemier, X. S. Hu, C. Zhuo, K. Ni, Fecam: A universal compact digital and analog content addressable memory using ferroelectric. *IEEE Trans. Electron Devices* **67**, 2785–2792 (2020).
22. C. Li, C. E. Graves, X. Sheng, D. Miller, M. Foltin, G. Pedretti, J. P. Strachan, Analog content-addressable memories with memristors. *Nat. Commun.* **11**, 1638 (2020).
23. C. Li, C. E. Graves, X. Sheng, D. Miller, M. Foltin, G. Pedretti, J. P. Strachan, A scalable design of multi-bit ferroelectric content addressable memory for data-centric computing, in *2020 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2020), pp. 29–33.
24. A. Kazemi, M. M. Sharifi, A. F. Laguna, F. Müller, R. Rajaei, R. Olivo, T. Kämpfe, M. Niemier, X. Sharon Hu, In-memory nearest neighbor search with FeFET multi-bit content-addressable memories. arXiv:2011.07095 [cs.ET] (2020).
25. Y. Xiao, Y. Xu, Z. Jiang, S. Deng, Z. Zhao, A. Mallick, L. Sun, R. Joshi, X. Li, N. Shukla, V. Narayanan, K. Ni, On the write schemes and efficiency of FeFET 1T NOR array for embedded nonvolatile memory and beyond, in *2022 International Electron Devices Meeting (IEDM)* (IEEE, 2022), pp. 13.6.1–13.6.4.
26. Z. Jiang, Z. Zhao, S. Deng, Y. Xiao, Y. Xu, H. Mulaosmanovic, S. Duenkel, S. Beyer, S. Meninger, M. Mohamed, R. Joshi, X. Gong, S. Kurinec, V. Narayanan, K. Ni, On the feasibility of 1T ferroelectric fet memory array. *IEEE Trans. Electron Devices* **69**, 6722–6730 (2022).
27. K. Ni, X. Li, J. A. Smith, M. Jerry, S. Datta, Write disturb in ferroelectric FETs and its implication for 1T-FeFET and memory arrays. *IEEE Electron Device Lett.* **39**, 1656–1659 (2018).
28. U. Schroeder, S. Slesazeck, H. Mulaosmanovic, T. Mikolajick, Nonvolatile field-effect transistors using ferroelectric-doped HfO_2 films, in *Ferroelectric-Gate Field Effect Transistor Memories. Topics in Applied Physics, vol 131*, B. E. Park, H. Ishiwara, M. Okuyama, S. Sakai, S. M. Yoon, Eds. (Springer, 2020).
29. M. Trentzsch, S. Flachowsky, R. Richter, J. Paul, B. Reimer, D. Utess, S. Jansen, H. Mulaosmanovic, S. Müller, S. Slesazeck, J. Ocker, M. Noack, J. Müller, P. Polakowski, J. Schreiter, S. Beyer, T. Mikolajick, B. Rice, A 28nm HKMG super low power embedded

NVM technology based on ferroelectric FETs, in *2016 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2016), pp. 11–15.
30. M. Lederer, T. Kämpfe, R. Olivo, D. Lehninger, C. Mart, S. Kirbach, T. Ali, P. Polakowski, L. Roy, K. Seidel, Local crystallographic phase detection and texture mapping in ferroelectric Zr doped HfO_2 films by transmission-EBSD. *Appl. Phys. Lett.* **115**, 222902 (2019).
31. X. Yin, F. Müller, Q. Huang, C. Li, M. Imani, Z. Yang, J. Cai, M. Lederer, R. Olivo, N. Laleni, S. Deng, Z. Zhao, Z. Shi, Y. Shi, C. Zhuo, T. Kämpfe, K. Ni, An ultracompact single-ferroelectric field-effect transistor binary and multibit associative search engine. *Adv. Intell. Syst.* **5**, 2200428 (2023).
32. M. Lederer, A. Reck, K. Mertens, R. Olivo, P. Bagul, A. Kia, B. Volkmann, T. Kämpfe, K. Seidel, L. M. Eng, Impact of the SiO_2 interface layer on the crystallographic texture of ferroelectric hafnium oxide. *Appl. Phys. Lett.* **118**, 012901 (2021).
33. H. Mulaosmanovic, J. Ocker, S. Müller, U. Schroeder, J. Müller, P. Polakowski, S. Flachowsky, R. van Bentum, T. Mikolajick, S. Slesazeck, Switching kinetics in nanoscale hafnium oxide based ferroelectric field-effect transistors. *ACS Appl. Mater. Interfaces* **9**, 3792–3798 (2017).
34. H. Mulaosmanovic, S. Dunkel, M. Trentzsch, S. Beyer, E. T. Breyer, T. Mikolajick, S. Slesazeck, Investigation of accumulative switching in ferroelectric FETs: Enabling universal modeling of the switching behavior. *IEEE Trans. Electron Devices* **67**, 5804–5809 (2020).
35. H. Bae, S. G. Nam, T. Moon, Y. Lee, S. Jo, D.-H. Choe, S. Kim, K.-H. Lee, J. Heo, Sub-ns polarization switching in 25nm Fe FinFET toward post CPU and spatial-energetic mapping of traps for enhanced endurance, in *2020 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2020), pp. 31–33.
36. M. Jerry, P.-Y. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu, S. Datta, Ferroelectric FET analog synapse for acceleration of deep neural network training, in *2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2017), pp. 6.2.1–6.2.4.
37. X. Sun, P. Wang, K. Ni, S. Datta, S. Yu, Exploiting hybrid precision for training and inference: A 2T-1FeFET based analog synaptic weight cell, in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2018), pp. 3.1.1–3.1. 4.
38. M. Halter, L. Bégon-Lours, V. Bragaglia, M. Sousa, B. J. Offrein, S. Abel, M. Luisier, J. Fompeyrine, Back-end, CMOS-compatible ferroelectric field-effect transistor for synaptic weights. *ACS Appl. Mater. Interfaces* **12**, 17725–17732 (2020).
39. K. Ni, M. Jerry, J. A. Smith, S. Datta, A circuit compatible accurate compact model for ferroelectric-FETs, in *2018 IEEE Symposium on VLSI Technology* (IEEE, 2018), pp. 131–132.
40. O. Guehairia, A. Ouamane, F. Dornaika, A. Taleb-Ahmed, Feature fusion via deep random forest for facial age estimation. *Neural Netw.* **130**, 238–252 (2020).
41. S. A. Roseline, S. Geetha, S., Kadry, Y. Nam, Intelligent vision-based malware detection and classification using deep random forest paradigm. *IEEE Access* **8**, 206303–206324 (2020).
42. X. Cao, R. Li, Y. Ge, B. Wu, L. Jiao, Densely connected deep random forest for hyperspectral imagery classification. *Int. J. Remote Sens.* **40**, 3606–3622 (2019).
43. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
44. C. Sapsanis, G. Georgoulas, A. Tzes, D. Lymberopoulos, Improving EMG based classification of basic hand movements using EMD, in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE, 2013), pp. 5754–5757.
45. H. Lee, R. Grosse, R. Ranganath, A. Y. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 609–616 (Association for Computing Machinery, 2009).
46. S. Beyer, S. Dünkel, M. Trentzsch, J. Müller, A. Hellmich, D. Utess, J. Paul, D. Kleimaier, J. Pellerin, S. Müller, J. Ocker, A. Benoist, H. Zhou, M. Mennenga, M. Schuster, F. Tassan, M. Noack, A. Pourkeramati, F. Müller, M. Lederer, T. Ali, R. Hoffmann, T. Kämpfe, K. Seidel, H. Mulaosmanovic, E. T. Breyer, T. Mikolajick, S. Slesazeck, FeFET: A versatile CMOS compatible device with game-changing potential, in *2020 IEEE International Memory Workshop (IMW)* (IEEE, 2020), pp. 1–4.
47. A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," thesis, Massachusetts Institute of Technology, Cambridge, MA (2009).
48. Q. Zhang, Y. Liao, X. Wang, T. Zhang, J. Feng, J. Deng, K. Shi, L. Chen, L. Feng, M. Ma, L. Xue, H. Hou, X. Dou, C. Yu, L. Ren, Y. Ding, Y. Chen, S. Wu, Z. Chen, H. Zhang, C. Zhuo, M. Tian, A deep learning framework for ^{18}F-FDG pet imaging diagnosis in pediatric patients with temporal lobe epilepsy. *Eur. J. Nucl. Med. Mol. Imaging* **48**, 2476–2485 (2021).
49. R. Vattikonda, W. Wang, Y. Cao, Modeling and minimization of pMOS NBTI effect for robust nanometer design, in *Proceedings of the 43rd annual Design Automation Conference* (2006), pp. 1047–1052.
50. M. Poremba, S. Mittal, D. Li, J. S. Vetter, Y. Xie, Destiny: A tool for modeling emerging 3D NVM and EDRAM caches, in *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (IEEE, 2015).
51. B. Van Essen, C. Macaraeg, M. Gokhale, R. Preger, Accelerating a random forest classifier: Multi-core, GP-GPU, or FPGA?, in *2012 IEEE 20th International Symposium on Field-Programmable Custom Computing Machines* (IEEE, 2012).
52. J. Ma, J. K. Saul, S. Savage, G. M. Voelker. Identifying suspicious urls: An application of large-scale online learning, in *Proceedings of the 26th Annual International Conference on Machine Learning* (Association for Computing Machinery, 2009), pp. 681–688.
53. K. J. Lee, G. Kim, J. Park, H. J. Yoo, A vocabulary forest object matching processor with 2.07 M-vector/s throughput and 13.3 nJ/vector per-vector energy for full-HD 60 fps video object recognition. *IEEE J. Solid-State Circuits* **50**, 1059–1069 (2015).
54. S. A. Nene, S. K. Nayar, H. Murase, Columbia object image library (coil-100) (1996).
55. T.-W. Chen, Y.-C. Su, K.-Y. Huang, Y.-M. Tsai, S.-Y. Chien, L.-G. Chen, Visual vocabulary processor based on binary tree architecture for real-time object recognition in full-HD resolution. *IEEE Trans. Very Large Scale Integr.* **20**, 2329–2332 (2012).
56. J. Zhang, Z. Wang, N. Verma, In-memory computation of a machine-learning classifier in a standard 6T SRAM array. *IEEE J. Solid-State Circuits* **52**, 915–924 (2017).
57. M. Kang, A multi-functional inmemory inference processor using a standard 6t sram array. *IEEE J. Solid-State Circuits* **53**, 642–655 (2018).
58. Center for Biological & Computational Learning (CBCL) at MIT, http://poggio-lab.mit.edu/codedatasets [accessed 4 June 2021].
59. S. K. Gonugondla, M. Kang, N. Shanbhag, A 42pJ/decision 3.12 TOPS/W robust in-memory machine learning classifier with on-chip training, in *2018 IEEE International Solid-State Circuits Conference-(ISSCC)* (IEEE, 2018), pp. 490–492.
60. M. Kang, S. K. Gonugondla, S. Lim, N. R. Shanbhag, 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).
61. R. A. Fisher, The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188 (1936).

Acknowledgments
Funding: M.N. was supported in part by ASCENT, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. **Author contributions:** X.Y. and K.N. proposed and supervised the project. C.L. and Q.H. performed the SPICE simulation. F.M., N.L., and T.K. conducted experimental characterization of ACAM cell and array. S.D. and Z.Z. performed single-device measurement. M.L. and T.K. conducted the EBSD characterization. A.F.L., Z.S., M.I., Y.S., M.N., X.S.H., and C.Z. performed the benchmarking and system evaluation. All authors contributed to writing of the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 16 September 2023
Accepted 1 May 2024
Published 5 June 2024
10.1126/sciadv.adk8471