

A Hybrid Edge Classifier: Combining TinyML-Optimised CNN with RRAM-CMOS ACAM for Energy-Efficient Inference

Kieran Woodward, Eiman Kanjo, Georgios Papandroulidakis, Shady Agwa, *Member, IEEE*,
and Themis Prodromakis, *Senior Member, IEEE*

Abstract—In recent years, the development of smart edge computing systems to process information locally is on the rise. Many near-sensor machine learning (ML) approaches have been implemented to introduce accurate and energy efficient template matching operations in resource-constrained edge sensing systems, such as wearables. To introduce novel solutions that can be viable for extreme edge cases, hybrid solutions combining conventional and emerging technologies have started to be proposed. Deep Neural Networks (DNN) optimised for edge application alongside new approaches of computing (both device and architecture -wise) could be a strong candidate in implementing edge ML solutions that aim at competitive accuracy classification while using a fraction of the power of conventional ML solutions. In this work, we are proposing a hybrid software-hardware edge classifier aimed at the extreme edge near-sensor systems. The classifier consists of two parts: (i) an optimised digital tinyML network, working as a front-end feature extractor, and (ii) a back-end RRAM-CMOS analogue content addressable memory (ACAM), working as a final stage template matching system. The combined hybrid system exhibits a competitive trade-off in accuracy versus energy metric with $E_{\text{front-end}} = 96.23nJ$ and $E_{\text{back-end}} = 1.45nJ$ for each classification operation compared with $78.06\mu\text{J}$ for the original teacher model, representing a 792-fold reduction, making it a viable solution for extreme edge applications.

Index Terms—tinyML optimisation, knowledge distillation, analogue memory-centric classifier, RRAM-based content addressable memory

I. INTRODUCTION

The proliferation of edge computing devices has created an increasing demand for efficient deployment of Deep Neural Networks (DNN) in resource-constrained environments. Deep neural networks have achieved remarkable success across

K. Woodward, is with the School of Computer Science, University of Nottingham, UK

E-mail: {kieran.woodward1}@nottingham.ac.uk

E. Kanjo, is with the Computing Department, Imperial College London, London, UK and the Department of Computer Science, Nottingham Trent University, UK

E-mail: {eiman.kanjo}@ntu.ac.uk

G. Papandroulidakis, S. Agwa, and T. Prodromakis are with the Centre for Electronics Frontiers (CEF), Institute for Integrated Micro and Nano Systems, University of Edinburgh, UK

E-mails: {gpapandr, shady.agwa, t.prodromakis}@ed.ac.uk

The authors acknowledge the support of the EPSRC FORTE Programme (Grant No. EP/R024642/2), the EPSRC ProSensing Project (Grant No. EP/Y030176/1) and the RAEng Chair in Emerging Technologies (Grant No. CiET1819/2/93). This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) through the Turing AI World Leading Researcher Fellowship: Somabotics: Creatively Embodying Artificial Intelligence [grant number EP/Z534808/1].

various applications such as image classification [1], natural language processing [2] and speech recognition [3] among others. However, their deployment on edge devices remains challenging due to two fundamental bottlenecks, the high computational complexity of fully connected layers and the substantial energy cost associated with memory access operations [4]. These challenges are particularly notable in the final classification layers of neural networks, where dense matrix operations require numerous memory accesses and floating-point computations. This has led to significant interest in developing methods to compress and optimise neural networks while maintaining acceptable performance levels [5], [6].

The traditional approach to addressing these challenges has focused on model compression techniques, including pruning, quantisation, and knowledge distillation [7], [8]. While these methods have shown success in reducing model size and computational requirements, they primarily optimise within the constraints of conventional digital architectures. Such approaches still rely on matrix multiplication operations and maintain the fundamental memory access patterns that contribute significantly to energy consumption. Recent work [9] has demonstrated that memory access operations can consume more energy than arithmetic operations in neural network inference, highlighting the limitations of purely algorithmic optimisation approaches.

Current hardware architectures, particularly in edge computing environments, face limitations in achieving energy efficiency and low-latency performance [10]. The bottlenecks often arise from the high energy consumption of memory access operations, the computational complexity of fully connected layers, and the reliance on conventional digital processing. In resource-constrained environments, the demand for local data processing necessitates both innovative model optimisations and novel hardware solutions. Emerging analogue hardware accelerators such as RRAM-based architectures, offer a promising avenue for mitigating these challenges by enabling parallel in-memory computation and reducing the energy costs associated with traditional digital architectures.

To address these challenges, we present an architecture and co-design methodology that combines both optimised DNNs alongside emerging analogue hardware accelerators. More specifically, our approach combines advanced DNN model compression techniques with RRAM-based template pattern matching in the analogue domain, replacing the computationally expensive and energy-intensive fully connected layers

with more energy efficient comparison operations. Through dynamic knowledge distillation, we transfer the capabilities of a complex neural network to a smaller model specifically designed for template generation. These templates are then used for classification through pattern matching rather than traditional matrix multiplication and activation functions.

This approach offers several key advantages over conventional solutions. First, it eliminates the need for expensive floating-point operations in the dense multi-level perceptron classification stage, replacing them with lower complexity parallel comparison operations. Second, the template-based matching process aligns with emerging hardware architectures optimised for parallel pattern matching, enabling more efficient implementation. Third, our quantisation scheme for template generation is specifically designed to minimise memory requirements while maintaining classification accuracy. Finally, by considering both hardware and software constraints in our design, we achieve a more holistic optimisation that addresses both computational complexity and energy efficiency.

In this work, we are showcasing a mixed-signal approach in designing an energy efficient edge classifier by integrating optimised DNN techniques with ACAM-based pattern matching.

- We propose a novel co-design methodology that combines digital DNN optimisation techniques with analogue hardware, enabling energy-efficient and low-latency edge classification.
- We introduce a knowledge distillation and pruning framework that compresses a pre-trained DNN into a compact model optimised for edge inference. This approach ensures high accuracy while reducing computational complexity, making it well-suited for deployment on resource-constrained devices.
- We develop robust techniques tailored for the generation of binary templates for ACAM-based classification and explore the impact of multi-template strategies and clustering methods on classification accuracy.

The remainder of this paper is organised as follows: Section II presents the model optimisation methodology, including knowledge distillation, pruning strategies, and quantisation schemes tailored for deployment on ACAM hardware. Section III introduces the ACAM hardware design, highlighting its memory-centric architecture and energy-efficient template matching capabilities, Section IV describes the experimental setup, Section V evaluates the results and finally Section VI concludes.

II. MODEL OPTIMISATION

In this section, we present our methodology for developing a compressed neural network model optimised for deployment on Analogue Content-Addressable Memory (ACAM) hardware. Our approach combines state-of-the-art model compression techniques with optimisations to create an efficient and edge computing hardware-friendly model [11].

Our proposed approach consists of several stages designed to contribute to the overall goal of creating an ACAM-optimised model:

- 1) Knowledge Distillation: A teacher-student framework is utilised, where a larger, more complex model (the teacher) is used to train a smaller, more compact model (the student). This technique enables the development of a model that maintains high accuracy while significantly reducing computational requirements and memory footprint.
- 2) Pruning: Following distillation, an iterative pruning strategy is applied to further reduce the model size. This step involves identifying and removing less important weights and connections, resulting in a sparser network structure.
- 3) quantisation: We implement a quantisation scheme that reduces the precision of weights and activations.
- 4) ACAM optimisations: Utilising the compressed model, we generate templates suitable for pattern matching, designed to utilise the analogue computation capabilities of ACAM hardware.

A. Knowledge distillation

Knowledge distillation forms a crucial component of the methodology for creating smaller, efficient models [12]. Our implementation follows a teacher-student framework where knowledge is transferred from a high-capacity teacher network to a more compact student network.

The knowledge distillation process is controlled by two key parameters: α and temperature T . The α parameter balances the importance between learning from the teacher's soft predictions and learning from the ground truth labels. The temperature parameter T controls the softness of probability distributions during knowledge transfer, where higher temperatures produce softer distributions that reveal more fine-grained knowledge about the relationships between the classes that the teacher has learned.

Traditional supervised learning typically uses *hard* labels for training. However, these hard labels don't capture the rich information about similarities between classes that the teacher model has learned. Knowledge distillation preserves this information by using the teacher's *soft* probability distributions as training input for the student model.

The knowledge distillation process is guided by a composite loss function that combines two components:

$$L = \alpha L_{KD}(z_s, z_t) + (1 - \alpha) L_{CE}(z_s, y) \quad (1)$$

where L_{KD} represents the knowledge distillation loss between the student outputs z_s and teacher outputs z_t , L_{CE} denotes the standard cross-entropy loss between student predictions and ground truth labels y , and α is the balancing parameter.

The knowledge distillation loss L_{KD} is computed using the Kullback-Leibler (KL) divergence between the softened probability distributions of the student and teacher models:

$$L_{KD}(z_s, z_t) = T^2 KL(\sigma(z_s/T) \parallel \sigma(z_t/T)) \quad (2)$$

where T is the temperature parameter that controls the softness of the probability distributions, and σ typically represents the activation function:

$$\sigma(z_i) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \quad (3)$$

The temperature scaling in equation 3 produces softer probability distributions by reducing the magnitudes of the logits, revealing more information about the inter-class relationships learned by the teacher. The T^2 scaling factor in equation 2 ensures the gradient magnitudes remain approximately constant when changing the temperature, as initially proposed by Hinton et al [12].

To improve the effectiveness of knowledge transfer, we incorporate curriculum learning by sorting the training data based on the teacher model's confidence. For each training sample (x_i, y_i) , we calculate a difficulty score d :

$$d(x_i, y_i) = L(z_t(x_i), y_i) \quad (4)$$

where $z_t(x_i)$ is the teacher's prediction for input x_i , and L represents the cross-entropy loss between the prediction and true label y_i . The training data is then ordered from easiest (lowest loss) to hardest (highest loss), allowing the student to gradually progress from simple to more challenging examples during training.

This knowledge distillation framework enables our student model to benefit from the rich representations learned by the larger teacher network while maintaining a significantly smaller parameter count suitable for edge deployment. The combination of soft targets and curriculum learning helps the student model learn more effectively, leading to better generalisation despite its reduced capacity.

B. Pruning strategy

After knowledge distillation, we apply structured pruning to further reduce the model size and computational requirements while maintaining performance. We employ a magnitude-based pruning approach that systematically removes less important weights based on their absolute values, followed by fine-tuning to improve accuracy [13].

The pruning schedule follows a polynomial decay that gradually increases network sparsity from an initial value $s_i = 0.50$ (50% sparsity) to a final target sparsity $s_f = 0.80$ (80% sparsity). This aggressive pruning strategy was chosen to substantially reduce model complexity while maintaining acceptable performance through the iterative pruning and fine-tuning process. The sparsity at step t is calculated as:

$$s(t) = s_f + (s_i - s_f)(1 - \frac{t}{n_t})^3 \quad (5)$$

At each pruning step, weights are ranked according to their absolute magnitude:

$$r(w_{ij}) = |w_{ij}| \quad (6)$$

where w_{ij} represents the weight connecting neurons i and j . Weights with the lowest magnitude are considered less important and are candidates for pruning. The pruning threshold θ_t at step t is determined by:

$$\theta_t = Q(|W|, s(t)) \quad (7)$$

where $Q(|W|, p)$ returns the p -th percentile of the absolute weights $|W|$, and $s(t)$ is the target sparsity at step t . Any weight with magnitude below this threshold is set to zero.

To maintain model performance, we implement an iterative pruning and fine-tuning schedule. After each pruning step, the network undergoes a brief period of fine-tuning to allow the remaining weights to adapt and compensate for the pruned connections. This process continues until the final target sparsity is reached.

After the pruning process is complete, a final fine-tuning phase is performed to ensure the pruned model maintains its accuracy. The remaining non-zero weights are then stored using a sparse matrix format, significantly reducing the model's memory footprint. The combination of gradual sparsity increase and iterative fine-tuning helps preserve the most important features learned by the network while eliminating redundant parameters. This compressed format is particularly well-suited for deployment on the edge, as it reduces both storage requirements and computational complexity during inference.

C. Quantisation scheme

Our quantisation strategy operates in two stages: first, quantizing the model weights using quantisation-aware training, and second, quantizing the feature maps to align with hardware constraints.

The model weights are quantised to 8-bit integers during training. This process applies quantisation during training, allowing the model to adapt to reduced precision, while still helping to preserve accuracy compared with lower bit quantisation during training. For ACAM deployment, we further quantise the feature maps. We use a mean-based thresholding approach for binary quantisation. This approach, using the mean rather than a fixed threshold, better adapts to the distribution of feature values.

D. ACAM-aware optimisations

Several architectural decisions in our network design are specifically guided by ACAM hardware constraints and efficiency considerations. The primary optimisation focuses on minimising the use of complex layer types and reducing the overall network depth while still aiming to maintain performance.

A key design decision is the exclusion of fully connected layers throughout the network. Traditional neural networks often employ multiple fully connected layers, particularly in their final stages, which significantly increases the parameter count and computational complexity. Instead, our architecture relies primarily on convolutional layers, which are more parameter-efficient due to weight sharing and maintain spatial relationships in the feature maps while also demonstrating superior memory access patterns. In convolutional operations, each loaded weight is reused multiple times as it slides across the input feature map and input values are reused across multiple

filter applications. This repeated use of data once loaded from memory significantly improves the energy efficiency of the network, as the energy cost of memory access is spread over multiple computations. This design choice not only reduces the model size but also aligns better with ACAM's pattern matching capabilities.

The number of convolutional layers is also carefully optimised. Rather than utilising deep networks with numerous convolutional layers, we employ a compact design with strategically placed convolutional layers. By using a progressive expansion and reduction in filter count we aim to maintain essential feature extraction capabilities while minimising the overall computational requirements. The resultant lightweight model is then used for the template generation and pattern matching that replaces the classification functionality of a traditional softmax layer.

1) Template generation: Template generation is a crucial step in optimising neural networks for deployment on ACAM hardware [14]. This process transforms the learned representations of a trained neural network into a format that can be efficiently stored and matched within the ACAM architecture. The primary goal of template generation is to create a set of representative patterns for each class that capture the features learned by the network.

Template generation aims to distill the high-dimensional feature representations learned by the neural network into a more compact and hardware-friendly format. This is achieved by analysing the logits of the network's penultimate layer (before SoftMax classification) when processing training samples from each class. These logits are then processed to create templates that capture the most important features for distinguishing between classes.

The trained neural network is used to generate feature maps for a large number of training samples from each class. These feature maps represent the network's internal representation of the input data.

Template generation is a crucial step in optimising neural network models for ACAM hardware implementation. This process involves converting the feature maps produced by the neural network into binary representations that can be efficiently stored and processed in ACAM architecture. Two primary methods for this binary thresholding have been explored; mean-based and median-based approaches.

In conventional signal processing, median-based thresholding is often preferred for its robustness to outliers and ability to maintain a balanced distribution of values in the binary templates [15]–[17]. However, our analysis (see Fig. 1) reveals that for neural network feature maps, mean-based thresholding consistently outperformed the median approach. In the mean-based method, the average value of each feature across all training samples is calculated to serve as the threshold, with feature values above the mean set to 1 and those below set to 0. While this approach is traditionally considered more sensitive to outliers, it can capture more nuanced distributions in the feature space which is advantageous in the context of neural network activations. Neural networks, particularly those employing ReLU activations, produce feature maps characterised by significant sparsity, with numerous zero values. These zero

activations substantially lower the mean threshold compared to the median, creating a more discriminative binary representation. The lower threshold enables the binary templates to capture subtle, non-zero activations that often carry important class-specific information, whereas a median-based approach might eliminate these features by setting a threshold that effectively treats many informative, low-magnitude activations as noise. Therefore, the mean-based method is more sensitive to the actual distribution of active features, preserving more relevant information for classification.

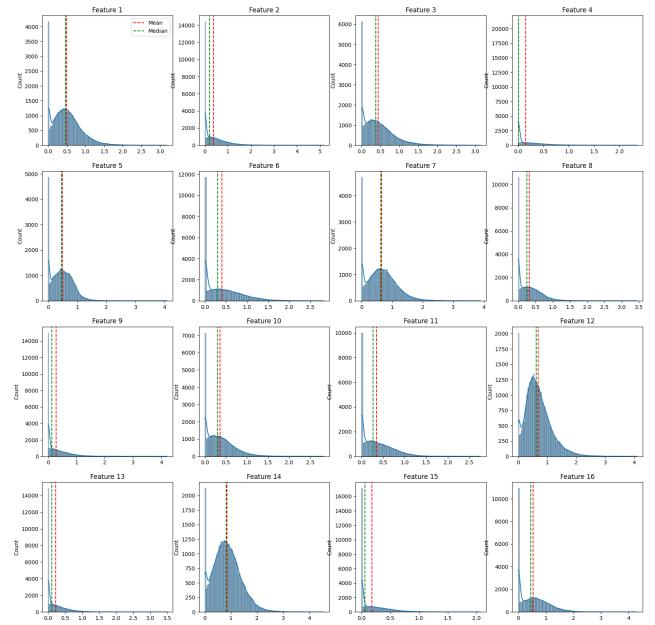


Fig. 1. Comparison of mean (red) and median (green) thresholding for each feature output from the front-end classifier.

We also explore the possibility of generating multiple templates for each class, specifically exploring configurations with one, two, or three templates per class. This approach acknowledges the variability within class distributions and aims to capture a more comprehensive representation of each class's feature space. By creating multiple templates, we can encompass different sub-clusters within a class, potentially leading to improved classification accuracy. The multi-template strategy is particularly beneficial for classes with high intra-class variability. In our implementation, these multiple templates are generated using a clustering approach on the feature representations of each class. For instance, when using three templates per class, we apply k-means clustering with $k=3$ on the class-specific feature maps, and the centroids of these clusters serve as the templates. K-means was chosen for its computational efficiency and ability to partition data into spherical clusters, making it a well-suited for this task. During the matching process, an input query is compared against all templates for each class, and the best match among these templates determines the class similarity score.

To further optimise the clustering process, we utilised silhouette scores to evaluate the quality of clustering for different configurations. Silhouette scores provided a quantitative measure of how well-separated and cohesive the clusters

are, enabling the selection of the number of templates that best captured the intra-class variability. This ensures that the generated templates represent meaningful and distinct sub-clusters within each class. While the multi-template approach increases the memory footprint and the number of comparisons required, it offers a trade-off between model complexity and representational power, potentially leading to more robust and accurate classifications, especially for complex datasets with significant intra-class variations.

2) *ACAM Pattern Matching Computing Concept*: We propose using pattern matching to serve as the inference mechanism, replacing the traditional final classification function of conventional digital DNNs with a hardware-optimised similarity search operation. The main principle of pattern matching is to compare an input pattern against a set of stored templates. In this work this is performed by a RRAM-CMOS ACAM system which is discussed in Section III. To evaluate the performance of the hybrid software-hardware classifier, the modelling of the ACAM is required towards introducing a good approximation of its behaviour to the software flow. Additionally, due to the intrinsic properties of the RRAM devices, the employment of program-once-read-many can provide better stability and endurance of the emerging RRAM devices. Thus, the calibration of the appropriate weights that are required for the RRAM-based ACAM in software and the programming once to the hardware is considered a more pragmatic method. For inference (assuming that the appropriate weights have been programmed to ACAM during training phase), the input pattern typically represents the feature map generated by processing an input sample through the front-end convolutional feature extractor network. This feature map is used as a query key to be compared against all stored templates. The stored templates have been generated during the training process to be used as the learned representation classes to be classified by the network. The comparison, which can be performed in parallel for all stored templates, identifies the template that most closely matches the input pattern, therefore determining the classification output by selecting the template that resulted in the highest similarity. A simple feature count pattern matching model is first explored to emulate the main functionality of ACAM without introducing the more hardware-demanding distance calculation from the matching window. Thus, unlike a more complex similarity calculation model, we are assuming that if an input is outside the effective matching window of each ACAM cell then its effect to the output is zero. Each input sample's quantised feature map is compared against the stored templates for all classes. The comparison sums the number of exactly matching features between the input and each template. This method is chosen for its simplicity and efficiency in hardware implementation, as it primarily involves bitwise comparison operations that can be performed efficiently. The class with the highest feature count is selected as the predicted class for the input sample.

For feature count-based pattern matching, given a query feature vector Q and template T , each with N features, the feature count score S_{fc} can be expressed as:

$$S_{fc}(Q, T) = \sum_{i=1}^N I(Q_i = T_i) \quad (8)$$

where I is the indicator function and ϵ is the matching threshold. For a more realistic model of ACAM, we are also testing a similarity-based pattern matching approach which implements a distance calculation when the input is outside the matching window of the ACAM cell. For each template, the function calculates a similarity score based on how well the query's features fall within the template's defined bounds. Specifically, it computes both a distance measure and a hit ratio. The distance is calculated as the sum of squared differences between the query and the template bounds, but only for features that fall outside the attribute's matching range. This approach allows for a degree of flexibility in pattern matching, accommodating small variations in feature values. The hit ratio represents the proportion of features that fall within the template's bounds, providing a measure of overall match quality. The final similarity score is derived from these metrics, with higher similarities indicating closer matches. This similarity-based approach provides a more granular assessment of match quality compared to the feature count method, potentially offering improved discrimination between closely related patterns.

For similarity-based pattern matching, we calculate both distance and hit ratio. Given a template T with upper and lower bounds $[T^L, T^U]$, the distance score D for features outside the template bounds is:

$$D(Q, T) = \sum_{i=1}^N \begin{cases} (Q_i - T_i^U)^2 & \text{if } Q_i > T_i^U \\ (T_i^L - Q_i)^2 & \text{if } Q_i < T_i^L \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

The hit ratio H measures the proportion of features falling within the template bounds:

$$H(Q, T) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(T_i^L \leq Q_i \leq T_i^U) \quad (10)$$

The final similarity score combines both metrics:

$$S_{sim}(Q, T) = \frac{H(Q, T)}{1 + D(Q, T)} \quad (11)$$

For both methods, the final classification decision C for a query Q across M classes is determined by:

$$C(Q) = \arg \max_{j \in \{1, \dots, M\}} S(Q, T_j) \quad (12)$$

where S represents either S_{fc} or S_{sim} depending on the chosen matching method.

III. RRAM-CMOS ACAM HARDWARE DESIGN AND OPERATION

In this section, we are introducing the hardware description of the RRAM-CMOS ACAM system used to replace the final dense layers in conventional DNNs. The specific technology employed for this ACAM is called Template piXeL (TXL)

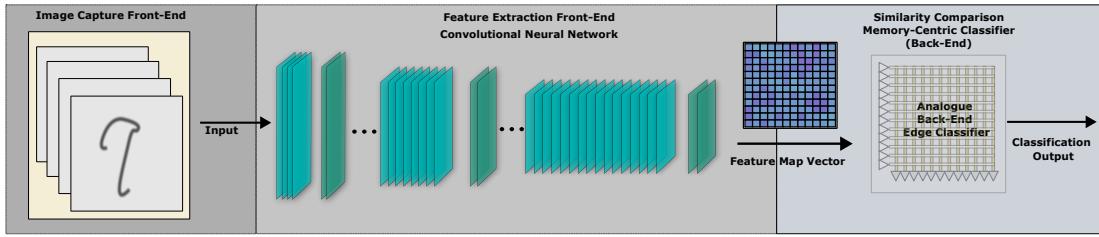


Fig. 2. Combined Deep Neural Network (DNN), for feature extraction and dimensionality reduction, with Analogue Content Addressable Memory (ACAM) back-end network [20], for final classification of the extracted feature map through analogue information processing.

and depending on the specific requirements of each edge application, different circuit and/or system implementations can be used. Each and every implementation discussed is used to implement specifically the last similarity comparison back-end system, as shown in Fig. 2.

A. ACAM Architecture and Background

Template matching hardware accelerators can be used as energy efficient classification engines closely integrated with edge sensors. Recently, content addressable memory (CAM) is a popular system architecture that has been repeatedly deployed as template matching accelerators [18]. CAMs are memory-centric systems using custom dense memory arrays with hardware-level optimisation aimed at performing a fast parallel search and match operation between a query key and the stored templates. CAMs operate in a massively parallel manner, comparing the input with every stored templates in the CAM array simultaneously [18]. In recent years, towards further reducing the energy dissipation of CAM systems, emerging memory technologies have been proposed as strong candidates to implement analogue memory and computing technologies. Some of these advances resulted in Analogue CAMs (ACAMs) that can be used to directly interface with analogue sensors' output and process information on-the-fly without analogue-to-digital conversion that is costly in terms of energy and IC area [19]. Implementing analogue template matching accelerator near the edge sensors can further help with data transfer bottlenecks since the information can be efficiently processed using real-valued data locally to the sensor and only the classification results will need to be transmitted.

On the device engineering front, recent advances in emerging memory technologies introduce a new component for novel analogue circuit design. Memristors, also known as Resistive Random Access Memory (RRAM), are two-terminal, tuneable, non-volatile, nanoscale resistive memory devices [21]. RRAM has many beneficial traits over conventional memory devices. RRAM can be integrated to implement novel computing systems showcasing low power consumption, high throughput, low area of integration and multi-bit information stored per cell compared to their fully CMOS counterparts.

In the last decade, RRAM has been extensively investigated for its use in processing information in the analogue domain while being organised in dense topologies. RRAM devices can be used to store multiple bits of information while retaining the

stored data in the absence of power, thus enabling dense non-volatile memory modules. The investigation of RRAM devices as a form of low power tuneable resistive element drives analogies with the observed behaviour of biological synaptic connections. Hence, novel DNNs using RRAM-CMOS hybrid circuits and systems have been proposed to satisfy the computing needs, especially for resource-constrained edge applications [22], [23]. Furthermore, the integration of RRAM devices into CAM architectures can result in novel energy-efficient ACAMs [24]. RRAM-CMOS ACAMs have been shown to be leveraged as an efficient component for ML models [25].

Towards integrating such analogue template matching classifiers in a wide variety of applications that require the classification of high-dimensional data, the use of a feature extractor front-end is required. A concept schematic of such a system is shown in Fig. 2. The feature extractor performs the necessary dimensionality reduction by generating a compressed feature map of the initial high-dimensional input. As an example, we could consider the use of an image sensor capable of capturing a still image with a specific number of sensing pixels with a convolutional network being used to perform a spatial feature extraction of the input. The flattened feature map (e.g. the output of a convolutional-based feature extractor) is used as a query input for the back-end RRAM-CMOS ACAM classifier. The classifier is used to perform the final classification of the lower-dimensional feature map by comparing it to its pre-stored feature map templates. The stored templates are created during the network training and are stored in ACAM for inference. The back-end classifier responds with a real-valued output vector that indicates the similarity of the feature map with each of the stored template/landmark feature maps, with the highest similarity being selected for the classified pattern. The similarity search is performed using real-valued data which circuit-wise is implemented by the analogue ACAM layer. Through this configuration of mixed-signal components, we can exploit the benefits of both digital and analogue information processing for ML applications. It has been shown that efficient digital feature extraction system can be implemented through tinyML techniques that can achieve competitive accuracy while using only a small fraction of resources compared to their full DNN counterparts [26]. Towards further optimising such techniques for extreme edge cases an analogue RRAM-based classifier can be used as final classification stage to replace the dense Multi-Layer Perceptron (MLP) network that conventionally is used for the

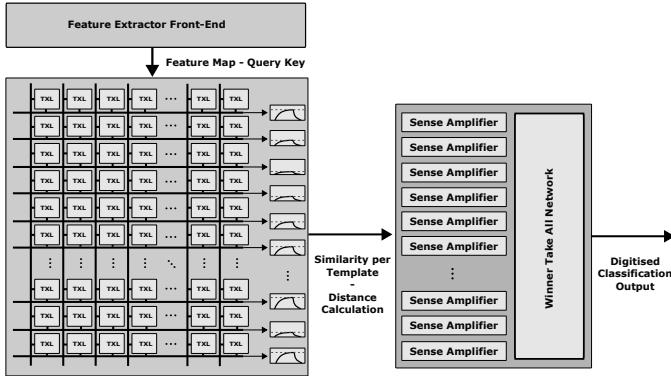


Fig. 3. Analogue Content Addressable Memory (ACAM) block diagram showcasing the main analogue computing blocks [28]. The block is effectively a two layer network, with the first layer being the ACAM module that calculates in parallel the similarity of the input feature maps with the pre-computed templates. The second layer senses the similarities and converts them into the proper voltage levels to pass through a final Winner Take All (WTA) network that computes the argmax function on the set of similarities. We assume that the feature map used as input to ACAM is the output of the front-end feature extractor with the ACAM being employed as a final layer classification network.

final fully connected layers in many DNN models. The back-end classifier can process data using the real-valued format of the feature map and can perform the similarity check in a massively parallel manner, thus saving both memory-computing resources as well as inference latency. Depending on the RRAM technology and the ACAM design employed, different weight quantisation formats can be used. The use of RRAM devices enables the integration of multi-bit memory technology near the computing, thus further saving memory resources and inference latency.

B. Memory-Centric Accelerators for Energy Efficient Template Matching

In this work, we are employing specifically Template piXeL (TXL) ACAM (TXL-ACAM) technology, which implements an energy-efficient ACAM system for accelerating near-sensor template matching operations. The ACAM system is designed in a commercially available 180 nm CMOS technology while the RRAM devices are added through back-end-of-line (BEOL) integration. The RRAM devices are designed and fabricated through in-house facilities [27]. The RRAM devices per cell are used to define the matching window, thus the voltage range for which an input is considered as matching. The TXL-ACAM implementation showcase competitive energy efficiency compared to other state-of-art ACAM technologies of approximately $185fJ$ per similarity search operation per cell.

The TXL-ACAM can classify a low dimensionality input (e.g. the feature map at the output of a convolutional feature extractor) by calculating the distance of the input vector to stored templates. A block diagram of the proposed TXL-ACAM is shown in Fig. 3. The output of the TXL-ACAM is a real-valued/analogue vector that encodes the similarity of the input per template. This computationally cornerstone operation could be considered similar to the search and match of a query

inside a miniaturised database. The set of similarities is processed through a Winner Take All (WTA) network to calculate efficiently the maximum similarity in the analogue domain. The WTA implements the argmax function and provides a one-hot encoding of the highest similarity template. All operations, with regard to the TXL-ACAM system, are implemented in the analogue domain using custom analogue circuitry. This results in an analogue back-end classifier that can replace a larger last stage CMOS-based MLP network used conventionally for the last few stages of DNNs.

TABLE I
TXL ACAM HARDWARE CHARACTERISTICS

Hardware Characteristics	RRAM-CMOS TXL ACAM [28]
CMOS Technology	180nm MOSFET (5V Components)
RRAM Technology	Pt/ AlO_{x-2} /Ti O_x /Pt-based metal oxide bi-layer MIM
Voltage Supply (Main)	3.3V
Energy (per cell per search)	$185fJ$
Frequency	10MHz
Latency (per inference/search)	100nsec
Area	$0.485m^2$
Macro Capacity	1536 cells 32×48 array

In Fig. 4(a), a schematic representation of the 6T4R TXL-ACAM pixel is shown [28]. The upper and lower thresholds of the matching window per cell are effectively defined based on the ratio of the upper and lower RRAM device which shifts the voltage threshold of the hybrid RRAM-CMOS inverters. Additional circuitry implemented per cell is used to conditionally charge the matchline when the input falls within the cell's matching window. An additional pMOS devices is controlling the cell's output through current limiting. The current limiter pMOS devices is used to calibrate the cell's charging rate in case of a match. The circuit performs a pattern matching operation between its input signal and its stored patterns (stored in the form of programmable conductances through the integration of the non-volatile RRAM devices). Each hybrid inverter stores one of the bounds through appropriately configured RRAM devices. The charge for each cell is accumulated using an analogue capacitor-based integrator circuit implemented for each row of the memory topology (thus per template). If multiple cells have a match with attributes of the query input then multiple connections to V_{DD} are enabled and the matchline is driven at some specific rate (dependent on the number of match enable available per TXL-ACAM array) to high voltage. The sense amplifiers can be used to detect if the overall match activity per template exceeds a specific predefined level controlled by the amplifiers threshold voltage. The sense amplifiers are calibrated to detect a specific voltage level which is translated to match based on the time-to-charge dynamics of the matchline. In case no such voltage level is observed within the readout operation timeframe, then the template is considered to not match with the input query at a sufficient rate. The sense amplifiers voltage threshold can be arbitrarily set depending on the intrinsic RRAM-CMOS cell dynamics that define the charging rate of the matchline.

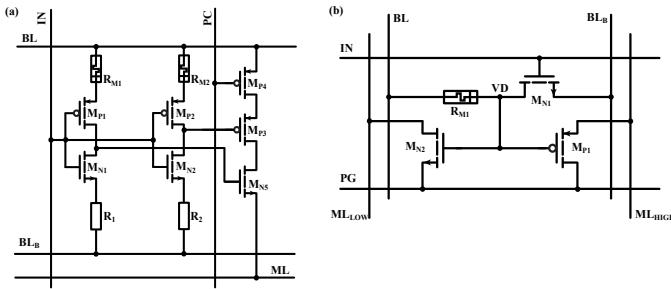


Fig. 4. RRAM-CMOS-based Template piXeL (TXL) ACAM cell schematics. There are many versions of the TXL-ACAM technologies with each version comprised of a specific set of trade-offs. In (a) a 6T4R charging design is showcased aimed at ML applications with increased sparsity [28]. In (b), a 3T1R precharging design is shown that is aimed at applications that have strict area specifications as well as differentiability as trait for the final stage classification [29].

In Fig. 4(b), a schematic representation of the 3T1R TXL-ACAM pixel is shown [29]. The 3T1R ACAM cell consists of a single 1-transistor-1-resistor (1T1R) memory cell (serial connection of R_{M1} RRAM and M_{N1} nMOS devices) and two additional transistors (M_{P1} pMOS and M_{N2} nMOS) for matchlines evaluation. The input voltage controls the resistance of the M_{N1} transistor, and thus the voltage divider formed by the RRAM devices and the M_{N1} . The additional nMOS and pMOS devices (M_{N1} and M_{P1} , respectively) are connected to a dual matchline configuration (ML_{LOW} and ML_{HIGH}). The two matchlines ML_{HIGH} and ML_{LOW} represent the high and low bounds of the matching window, respectively. Due to the complementary behaviour between nMOS and pMOS, the nMOS transistor is responsible for discharging ML_{LOW} when the input voltage is below the lower bound while the pMOS transistor is responsible for discharging ML_{HIGH} when the input voltage exceeds the higher bound. Contrary to Fig. 4(a), the 3T1R design follows a precharging and evaluating operation cycle instead of a discharging/initialisation and evaluation. The larger charging cell could be preferable in sparse activation applications, while the more conventional 3T1R precharging cell design could be preferred in normally distributed activation applications due to its smaller size.

More specifically, if the input voltage IN is below the lower bound V_{LOW} , the resistance of the nMOS transistor M_A is relatively high which drives VD (voltage at the intermediate node of the 1T1R voltage divider circuit) to an appropriate voltage to enable discharge through the nMOS transistor M_{MLO} of matchline ML_{LOW} . At the same time, the pMOS transistor M_{MHI} is non-conductive and the matchline ML_{HIGH} is not discharging through the pMOS device. If the input voltage IN exceeds the higher threshold, the resistance of the nMOS transistor M_A is relatively low which drives VD to an appropriate voltage to make the pMOS transistor M_{MHI} ON to discharge the matchline ML_{HIGH} while the nMOS transistor M_{MLO} is OFF and the matchline ML_{LOW} is not discharging fast enough. If the input voltage IN is lower than the high threshold and higher than the low threshold, then a match case occurs. The safe VD value makes both pMOS transistor M_{MHI} and nMOS transistor M_{MLO} partially

or completely OFF so that the matchlines ML_{HIGH} and ML_{LOW} are not discharging. By evaluating both matchlines a final decision about the match/mismatch case can be asserted. Furthermore, being able to evaluate each threshold separately provides better assessment in case of mismatching and makes the cell differentiable. Thus, we can understand how to train the RRAM weights by assessing which threshold is surpassed or not.

Due to the use of analogue sub-systems the noise sensitivity is an area that requires considerations to avoid non-idealities from the analogue and mixed-signal sub-system to introduce noise to the final classification output. The TXL ACAM system employs supra-threshold RRAM-CMOS operating point thus mitigating the noise introduced by the CMOS circuits to the final classification results. The main noise source could be traced to the emerging RRAM technology due to cycle-to-cycle and device-to-device variability as well as other non-idealities. The functional model of the TXL ACAM used to implement the proposed model does not model the hardware non-idealities in the Python simulations provided. This enables us to easily integrate this functional model of TXL to TensorFlow framework and test its extrapolated effectiveness in tinyML classification applications. The energy figure used for the model-level energy calculations is based on post-layout simulations and is used to calculate the approximate extrapolated performance of the custom tinyML network under investigation. The TXL ACAM layer has been simulated on Cadence Virtuoso environment and Spectre+SPICE simulation engine. Although RRAM devices have cycle-to-cycle variations, statistical corner analysis can help to define safety margins for the matching windows per each cell. This marginally affects the tuneability of the TXL cells which in return affects the quantization. However, the proposed 9T4R TXL design demonstrates high matching window tuneability which can be used to mitigate this variability by applying wider safety margins without significant effect on the quantization and approximation template matching computing.

IV. EXPERIMENTAL SETUP

A. Datasets preparation

The CIFAR-10 dataset [30] has been used to evaluate model performance as it is a widely recognised benchmark in the field of image classification. CIFAR-10 is selected as our evaluation benchmark because it represents a well-established standard in the TinyML community, enabling direct comparison with existing edge classification approaches while demonstrating the core hybrid co-design methodology. The grayscale preprocessing reflects realistic constraints in ultra-low-power edge vision systems where monochrome sensors are commonly employed to minimise data bandwidth and energy consumption. This dataset consists of 60,000 32x32 colour images across 10 classes, with 6,000 images per class. The dataset is divided into 50,000 training images and 10,000 test images.

CIFAR-10 serves as a common challenge in image classification tasks, allowing for direct comparison with other state-of-the-art methods. The dataset's size allows for rapid

experimentation and iteration, which is crucial when developing new classification approaches. Furthermore, the 10 distinct classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck) provide a good representation of real-world object recognition tasks.

To better simulate the input characteristics of ACAM hardware we pre-processed the CIFAR-10 images by converting them to grayscale. This conversion was performed using the standard formula: $Y = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B$, where Y is the resulting grayscale value, and R , G , and B are the red, green, and blue channel values respectively. This conversion is specifically motivated by ACAM hardware constraints, as reducing from 3-channel RGB to single-channel grayscale decreases memory requirements by 67% and simplifies analogue template matching operations in our RRAM-based cells. After the grayscale conversion, the values are normalised to improve the stability of neural network training and the subsequent quantisation process.

In addition to CIFAR-10, we validate our approach on a larger, more complex dataset, CINIC-10 [31]. CINIC-10 presents a significantly more demanding benchmark, comprising 270,000 images with greater diversity due to the inclusion of downsampled ImageNet samples, which introduces substantial domain shift and intra-class variability.

B. tinyML Feature Extractor Model architectures

A ResNet-50 [32] architecture is used as the teacher model. This deep residual network is chosen for its high performance on the CIFAR-10 dataset and its ability to learn rich, hierarchical features [33]. The ResNet-50 model consists of three stages, each containing multiple residual blocks. The first stage operates on 16 channels, with subsequent stages doubling the number of channels while reducing spatial dimensions. Each residual block comprises of two 3×3 convolutional layers with batch normalisation and ReLU activation function. A shortcut connection is added to each block, with 1×1 convolutions used when necessary to match dimensions. The network concludes with global average pooling and a fully connected layer for classification. The ResNet-50 model is initialised with He normal initialisation and employs L2 regularisation to prevent overfitting.

1) *Student Ablation Studies*: Our ablation studies quantify the impact of each optimisation technique and architectural decision on classification performance of the student model as shown in Table II. The study explored a range of architectures, from simple dense networks (390,000 parameters) to complex multi-layer convolutional models (5.9M parameters), exploring various configurations including dense layers (128-1024 units), convolutional layers (16-512 filters), and global average pooling. While not every configuration explored during this exploratory phase is shown, Table II highlights the key architectures, parameters and optimisation techniques we were considering before finalising the architecture choice. The sequential application of optimisation techniques demonstrated clear benefits. Knowledge distillation consistently improved accuracy across all model configurations, with an average improvement of 5.2% compared to baseline training. Notably, for the CNN-based architectures, knowledge distillation yielded even more

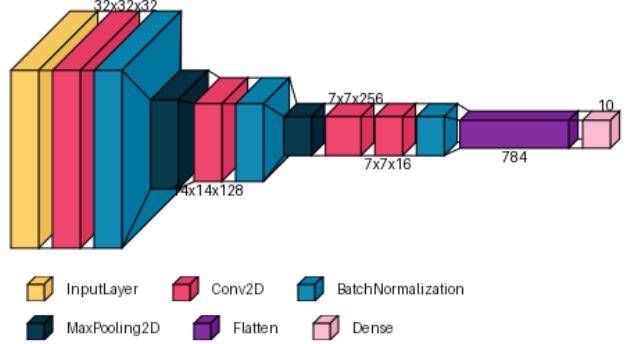


Fig. 5. Student CNN model architecture using a traditional softmax classifier.

substantial gains, with improvements ranging from 7.3% to 9.4%. The subsequent application of pruning maintained this enhanced performance, showing minimal degradation (average -0.8%) while significantly reducing model complexity. 8-bit training aware quantisation also proved extremely robust, reducing accuracy by only 0.42% on average while enabling hardware-efficient implementation.

Starting with a baseline model containing dense layers (Dense 128), we observed that increasing dense layer width (256, 512) actually degraded performance, with accuracies dropping to 78.2% and 77.89% respectively for binary similarity-based matching. This suggests that larger dense layers may introduce redundant parameters without capturing additional discriminative features. The final architecture, avoiding dense layers to align with ACAM hardware constraints, achieved the optimal balance between model complexity and accuracy with convolutional termination layers outperformed dense layer configurations in terms of stability across quantisation schemes. Models concluding with convolutional layers demonstrated consistent performance across different quantisation levels, with accuracy variations limited to $\pm 1.2\%$. In contrast, dense layer configurations showed higher sensitivity to quantisation, with accuracy variations of up to $\pm 3.5\%$.

2) *Student Model Architecture*: The final student model is a significantly smaller convolutional neural network designed with ACAM deployment as shown in Fig. 5. It consists of three main convolutional layers, where the first two are followed by batch normalisation and max pooling operations. The network begins with 32 filters in the first layer, increases to 128 in the second, and concludes with 256 filters in the third layer. An additional convolutional layer with 16 filters is appended to reduce the feature map size. The architecture has been designed to balance computational efficiency with model size and feature extraction capability.

To quantify the computational requirements of both architectures, we calculate the number of multiply-accumulate (MAC) operations. For each convolutional layer, the total number of MAC operations is given by:

$$MAC_{\text{conv}} = H_{\text{out}} \times W_{\text{out}} \times K_h \times K_w \times C_{\text{in}} \times C_{\text{out}} \quad (13)$$

where H_{out} and W_{out} are the output feature map dimensions,

TABLE II

ABLATION STUDY RESULTS: IMPACT OF KNOWLEDGE DISTILLATION (KD), PRUNING, AND QUANTIZATION ON DIFFERENT STUDENT MODEL ARCHITECTURES USING A SOFTMAX LAYER FOR CLASSIFICATION.

Architecture Description	Baseline Accuracy (%)	KD Accuracy (%)	Pruning Accuracy (%)	Quantization Accuracy (%)	Parameters	Model Size (MB)	Improvement (%)
Dense Layer Architectures							
Dense 128	41.91	45.59	38.49	43.78	390K	0.38	+3.68
Dense 256-128-64	46.97	50.19	48.62	49.52	800K	0.80	+3.22
Dense 1024-512-256	46.60	51.63	50.89	50.07	3.8M	3.64	+5.03
Dense 512-128	47.40	50.56	49.19	50.91	1.6M	1.57	+3.16
Simple CNN Architectures							
CNN 16 + Dense 32	46.16	45.22	43.68	47.03	460K	0.45	+0.87
CNN 32 + Dense 32	55.21	59.20	61.09	58.87	920K	0.88	+5.88
CNN 32 + Dense 64-16	57.92	63.10	63.92	60.83	1.8M	1.76	+6.00
Multi-Layer CNN Architectures							
CNN 16-32-64 + Dense 32	61.10	71.23	70.61	65.24	1.4M	1.35	+10.13
CNN 64-128 + Dense 32	59.83	71.35	71.64	66.30	3.3M	3.14	+11.81
CNN 32 + Dense 128	58.00	70.56	70.00	67.00	3.6M	3.50	+12.56
CNN 32-64 + Dense 64	59.36	71.89	71.47	67.26	3.2M	3.09	+12.53
CNN 16-32-64 + Dense 128	62.58	74.56	73.69	67.84	5.6M	5.31	+11.98
CNN 16-32 + Dense 32	65.77	73.36	72.85	67.89	3.2M	3.09	+7.59
CNN 64-128-256 + Dense 32	59.39	75.73	74.64	68.93	5.9M	5.65	+16.34
Global Average Pooling (GAP) Architectures							
CNN 64-128-256 + GAP	60.19	66.57	70.67	71.14	380K	0.38	+10.95
CNN 64-128-512 + GAP	65.37	71.28	75.44	76.05	832K	0.82	+10.68
CNN 64-128-256-1024 + GAP	70.85	78.16	79.61	77.74	3.0M	2.93	+8.76

CNN notation: Conv2D filters (kernel size), GAP: GlobalAveragePooling2D

Improvement column shows the best technique improvement over baseline

All models evaluated on grayscale CIFAR-10 dataset

K_h and K_w are the kernel dimensions, and C_{in} and C_{out} are the input and output channels respectively. This calculation allows us to compare the computational complexity of our teacher and student models, demonstrating the efficiency gains achieved through our architectural choices.

V. RESULTS

A. Model compression performance

The implementation of our compression methodology demonstrates a successful balance between model size reduction and maintained performance. As shown in Table III, our teacher model, based on the ResNet-50 architecture, achieved a baseline accuracy of 93.77% on the CIFAR-10 dataset, with corresponding F1-score, precision, and recall metrics all around 93%, which is similar to previous work [33]. This performance comes at the cost of substantial computational requirements, with the model containing approximately 26.2 million parameters and requiring roughly 3 billion operations. The same teacher model trained on the greyscale CIFAR-10 dataset requires a similar number of parameters but achieves a slightly reduced accuracy of 91.04% due to the reduced number of features. This demonstrates the increased challenge of working with greyscale images for the compressed student models.

On the other hand, the unoptimised base student model, while achieving a significant reduction in model size to just 380,314 parameters (a 98.5% reduction), initially showed a considerable performance drop to 76.29% accuracy. This model required nearly 24 million operations, already representing a 99% reduction in computational complexity compared to the teacher model. However, the application of our optimisation techniques, including knowledge distillation, pruning

and quantisation significantly improved the student model's performance while also further reducing the number of operations. The optimised student model achieved an accuracy of 82.22%, representing a substantial 5.93% improvement over its unoptimised comparative model. This optimisation came with an additional benefit of reducing the effective MAC operations to 4.76 million through 80% sparsity, as operations involving pruned (zero-valued) weights can be skipped entirely, representing an 800-fold reduction in operations compared to the original teacher model. By eliminating these unnecessary computations, the network achieves significant energy savings while maintaining classification accuracy.

While the performance gap between the teacher and optimised student model remains at 8.82%, it should be considered in the context of the dramatic reduction in model complexity. The optimised student model maintains this performance level while using just 1.45% of the parameters of the teacher model and requiring only 0.13% of the computational operations. This trade-off between performance and resource efficiency positions the model favourably for deployment on resource-constrained ACAM hardware.

B. Pattern matching performance analysis

The evaluation of different pattern matching approaches reveals important insights about the trade-offs between accuracy and computational complexity. Binary (1-bit) template quantisation achieved consistent performance of 70.91% accuracy across both feature count (see Fig. 6) and similarity-based approaches, though this represents a reduction from the 82.22% accuracy achieved by the student model using a traditional softmax classifier. This 11% accuracy drop must be considered in the context of the significant computational

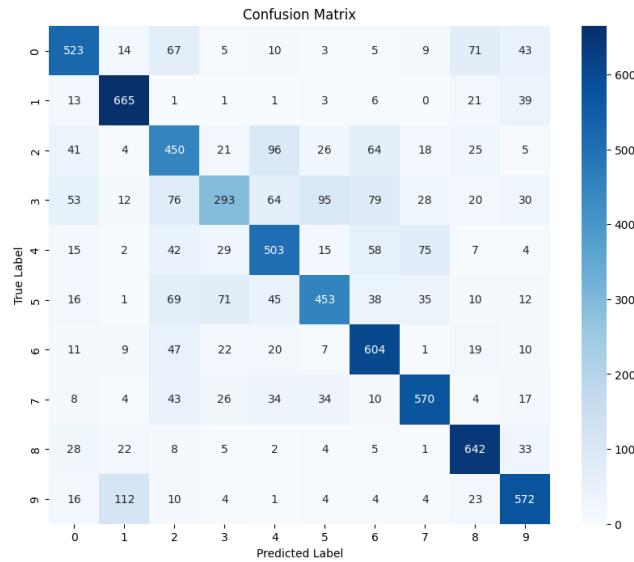


Fig. 6. Confusion matrix of optimised student model using feature-based pattern matching classifier.

benefits - while the softmax classifier requires floating-point operations and dense matrix multiplications, the binary pattern matching approaches operate using only simple bit comparisons.

While the 20% accuracy reduction from teacher model to final pattern matching implementation appears substantial, this trade-off becomes acceptable in several critical edge computing scenarios where energy efficiency and computational constraints outweigh absolute accuracy requirements. Applications such as always-on wearable health monitoring devices, where battery life is paramount and approximate classification is sufficient for trend detection, would benefit significantly from the energy reduction achieved by our hybrid approach. Similarly, resource-constrained IoT sensor networks for environmental monitoring, smart agriculture, or industrial predictive maintenance can tolerate moderate accuracy losses when the alternative is frequent battery replacement or inability to deploy due to power constraints. Additionally, applications requiring real-time response with minimal latency, such as basic gesture recognition for human-computer interaction or simple object detection for robotics, can leverage the computational simplicity of bit-wise operations over complex floating-point calculations, making the accuracy trade-off worthwhile for maintaining responsive performance on severely constrained hardware platforms.

The identical performance between feature count and similarity-based approaches in the binary domain can be explained by examining how the matching operations converge when features are quantised to single bits. In the binary case, both methods effectively reduce to counting matching bit positions, since the similarity approach's distance calculations and hit ratio measurements become equivalent to direct bit comparisons when working with binary values. This convergent behaviour suggests that more complex similarity calculations provide no additional benefit when working with these features.

While binary pattern matching achieves lower accuracy compared to the student model with softmax, its computational simplicity and minimal memory requirements make it an attractive option for extremely resource-constrained edge applications where the 11% accuracy trade-off is acceptable in exchange for significant energy savings and hardware simplification.

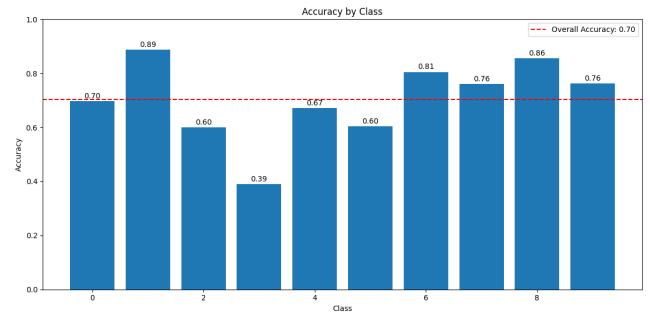


Fig. 7. Per-class accuracy of optimised student model using feature-based pattern matching classifier.

C. Multiple template analysis

The investigation of using multiple templates per class reveals slight improvements in classification accuracy, as shown in Table IV. This analysis explores configurations using one, two, and three templates per class to capture intra-class variations more effectively.

The increasing the number of templates from one to two improved accuracy from 70.91% to 71.64%, a gain of 0.73%. However, further increasing to three templates per class showed a slight decrease to 71.60%, suggesting diminishing returns from additional templates. This pattern suggests that binary quantisation, with its reduction of features, creates relatively distinct and well-separated class representations where a single template can already capture the essential pattern for each class. The minimal improvement with a second template and decrease with a third indicates that these features lack the granularity to meaningfully distinguish additional intra-class patterns, essentially leading to redundant or potentially conflicting templates.

These consistently small improvements indicate that increasing the number of templates may not be an effective strategy for enhancing classification performance. The limited gains must be weighed against the significant increase in memory requirements and computational complexity associated with storing and comparing multiple templates per class. This suggests that simpler configurations with fewer templates might offer a better balance between resource utilisation and classification performance. The results indicate that focusing on optimising the quality of a single template per class might be more beneficial than increasing the number of templates, particularly given the minimal returns observed with additional templates.

TABLE III
COMPARISON OF TEACHER AND STUDENT MODEL PERFORMANCE, CALCULATED USING A SOFTMAX LAYER FOR CLASSIFICATION.

Model	Accuracy	F1-Score	Precision	Recall	Parameters	MAC operations	Compression ratio
Teacher colour	93.77	93.78	93.83	93.77	26,215,810	3,858,551,808	-
Teacher greyscale	91.04	91.05	91.13	91.05	26,209,538	3,808,375,808	1.01:1
Student (without optimisations)	76.29	76.29	76.14	79.29	380,314	23,785,120	162:1
Student (with optimisations)	82.22	82.20	82.34	82.23	380,314	4,757,024	811:1

TABLE IV
COMPARISON OF ACCURACY WHEN USING VARYING NUMBER OF TEMPLATES.

Number of templates	Accuracy (%)
1	70.91
2	71.64
3	71.60

D. Energy efficiency estimates

For the back-end classifier, each template matching operation consumes 185fJ per cell. With our architecture requiring 10 templates of 784 features each, the energy consumption is:

$$E_{\text{back-end}} = N_{\text{templates}} \times N_{\text{features}} \times E_{\text{cell}} \\ = 10 \times 784 \times 185\text{fJ} = 1.45\text{nJ} \quad (14)$$

The total energy consumption of our front end network can be calculated considering both sparsity and quantisation effects. Given our total MAC operations of 23,785,120 from Table III, and accounting for 80% sparsity, the effective number of MAC operations is 4,757,024. However, we must also consider that lack of the final softmax fully-connected layer using this ACAM optimised approach, therefore we can remove an addition 7,850 operations that were used by this final layer, meaning the total operations for the front end classifier is 4,749,174. Using the energy figures from Horowitz [34] for 8-bit operations (0.2pJ for multiply and 0.03pJ for add) and memory access costs (20pJ for 32KB cache), we can calculate the energy consumption of the front-end feature extractor. For each MAC operation, the computation energy is 0.23pJ and the memory access energy is 20pJ, giving a total front-end energy consumption of 96.07nJ per inference. This combines with our back-end ACAM energy of 1.45nJ for template matching, resulting in a total system energy consumption of 97.52nJ per classification operation. In comparison the teacher model has a total energy consumption of 78.06μJ demonstrating the proposed methodology achieved a 792 times energy reduction.

This demonstrates that our hybrid approach achieves significant energy efficiency, with the ACAM back-end consuming less energy than traditional digital implementations. This paves the way for new opportunities in multi-modal sensing enabling smaller, more efficient deep learning models that can be implemented in devices such as wearables while preserving battery life.

E. Discussion

When compared to state-of-the-art digital-only TinyML optimisation techniques, our hybrid approach demonstrates

strengths that address specific limitations of purely digital methods. Recent digital TinyML frameworks such as MCUNet [35] and optimised MobileNetV3 implementations achieve compression ratios of around 7× while maintaining 80-90% accuracy on CIFAR-10 outperforming our 70.91% accuracy. However, these approaches still rely fundamentally on digital multiply-accumulate (MAC) operations and dense matrix multiplications in their classification layers which typically consume 200-2500 μJ per inference even after quantisation and pruning [36]. In contrast, our hybrid approach eliminates floating-point calculations entirely in the final classification stage, replacing computationally expensive softmax operations with simple bit-wise pattern matching that consumes only 1.45nJ for the back-end classifier. While binarized neural networks (BNNs) [37] achieve similar bit-level operations throughout the entire network with 88-90% CIFAR-10 accuracy, they still rely on sequential layer-by-layer processing where each layer must complete before the next can begin. This fundamental architecture prevents BNNs from exploiting the massively parallel template matching capabilities inherent in ACAM hardware, where all stored templates are compared simultaneously against the input query in a single operation cycle.

To address the scalability and generalisability of our proposed methodology beyond CIFAR-10, we conducted additional experiments on the more challenging CINIC-10 dataset. Training on grayscale CINIC-10, our teacher model (ResNet-56) achieved 72.84% test accuracy, while the optimised student model reached 65.96% accuracy. These results align well with literature our student model performance is comparable to lightweight architectures achieving 65-66% on CINIC-10 [38], while our teacher model performs competitively to state-of-the-art models which have achieved 82% [39], [40], considering the additional constraints of grayscale conversion. Following 1-bit feature-based template pattern matching using the optimised student model, the hybrid system maintained 53.74% accuracy. Despite the 12.22% accuracy reduction from student to template matching, the system maintains the energy reduction demonstrated on CIFAR-10, validating that our methodology generalises effectively to larger, more diverse datasets.

The fundamental advantage of our approach lies not in competing directly with digital optimisation techniques, but in addressing the memory wall problem that limits all digital approaches especially in edge computing case scenarios. Traditional TinyML methods, regardless of quantisation level, suffer from the fundamental energy cost imbalance between computation and memory access. According to Horowitz [34], memory accesses consume 10-100x more energy than

arithmetic operations (10-100 pJ for cache access vs. 0.03-0.9 pJ for basic operations), with off-chip DRAM accesses requiring 1-2 orders of magnitude more energy still (1.3-2.6 nJ). In conventional processors, on-chip memory systems already account for over 50% of total die energy consumption, and this proportion increases significantly when off-chip memory accesses are included. TinyML inference exacerbates this problem through repeated loading of weights and activations from memory for each layer's computations, creating a memory-bound rather than compute-bound energy profile. This memory access bottleneck remains problematic regardless of weight quantisation techniques, as even 1-bit weights must still be retrieved from memory storage during inference.

Our RRAM-based ACAM performs computation directly within the memory array, eliminating data movement between memory and processing units for the classification stage. This memory-centric computing paradigm becomes increasingly valuable as model sizes grow and memory access costs increase. Furthermore, the analogue nature of our back-end classifier enables natural handling of sensor noise and variations without requiring additional digital processing steps, making it particularly suitable for direct sensor-to-decision applications in IoT environments. While purely digital approaches excel in applications requiring maximum accuracy, our hybrid methodology offers a unique solution space for ultra-low-power applications where the combination of acceptable accuracy, extreme energy efficiency (97.52nJ total), and hardware simplicity creates deployment opportunities that would be infeasible with conventional digital-only TinyML approaches.

VI. CONCLUSIONS

This work demonstrates the feasibility of a hybrid digital-analogue architecture for edge image classification through the integration of optimised neural networks with RRAM-based ACAM hardware accelerators. The proposed digital methodology utilising knowledge distillation achieves substantial efficiency improvements while maintaining accuracy, with our resultant student model achieving 82.22% accuracy on CIFAR-10 using only 1.45% of the teacher model's parameters and reducing MAC operations by a factor of 800. For analogue classification, template pattern matching maintains 70.91% accuracy using a simple feature count approach, suggesting more complex pattern matching approaches may not yield proportional benefits in classification accuracy.

This approach demonstrates significant advantages in terms of energy efficiency, with the ACAM back-end consuming only 1.45nJ per classification operation compared with 78.06 μ J for the teacher model, while maintaining acceptable accuracy levels for edge applications. These results establish the potential of digital-analogue approaches for resource-constrained edge devices. Future research could further explore template generation techniques and novel ACAM cell designs, further optimising classification accuracy and energy efficiency in edge computing applications.

REFERENCES

- [1] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Computation*, vol. 29, no. 9, pp. 2352–2449, sep 2017.
- [2] D. W. Otter, J. R. Medina, and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 604–624, feb 2021.
- [3] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, p. 101869, nov 2023.
- [4] D. Nagaraju and N. Chandrachoodan, "Compressing fully connected layers of deep neural networks using permuted features," *IET Computers and Digital Techniques*, vol. 17, no. 3-4, pp. 149–161, jul 2023.
- [5] B. Zandonati, G. Bucagu, A. A. Pol, M. Pierini, O. Sirkin, and T. Kopetz, "Towards Optimal Compression: Joint Pruning and Quantization," feb 2023.
- [6] H. Li and L. Meng, "Hardware-aware approach to deep neural network optimization," *Neurocomputing*, vol. 559, p. 126808, nov 2023.
- [7] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, oct 2021.
- [8] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, jun 2021.
- [9] Y. Zheng, H. Yang, Y. Shu, Y. Jia, and Z. Huang, "Optimizing Off-Chip Memory Access for Deep Neural Network Accelerator," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 4, pp. 2316–2320, apr 2022.
- [10] F. Golpayegani, N. Chen, N. Afraz, E. Gyamfi, A. Malekjafarian, D. Schäfer, and C. Krupitzer, "Adaptation in Edge Computing: A review on design principles and research challenges," *ACM Transactions on Autonomous and Adaptive Systems*, sep 2024.
- [11] A. Mukherjee and G. Kulkarni, "Tinyml techniques for running machine learning models on edge devices," *Proceedings of the Second International Conference on AI-ML Systems*, 10 2022.
- [12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," 3 2015.
- [13] G. Yang, G. Hao, L. Weijia, W. Qinghua, S. Chen, and N. Zhang, "An Attentive Pruning Method for Edge Computing," *ACM International Conference Proceeding Series*, pp. 6–10, feb 2020.
- [14] F. Aguirre, A. Sebastian, M. Le Gallo, W. Song, T. Wang, J. J. Yang, W. Lu, M. F. Chang, D. Ielmini, Y. Yang, A. Mehonik, A. Kenyon, M. A. Villena, J. B. Roldán, Y. Wu, H. H. Hsu, N. Raghavan, J. Suñé, E. Miranda, A. Eltawil, G. Setti, K. Smagulova, K. N. Salama, O. Krestinskaya, X. Yan, K. W. Ang, S. Jain, S. Li, O. Alharbi, S. Pazos, and M. Lanza, "Hardware implementation of memristor-based artificial neural networks," *Nature Communications* 2024 15:1, vol. 15, no. 1, pp. 1–40, mar 2024.
- [15] M. Kantharimuthu, P. Selvaraj, H. Sankar, and G. Brindavanam, "Efficient Parallel Median Filter for Image Denoising: Implementation and Performance Evaluation," *Traitement du Signal*, vol. 41, no. 05, pp. 2403–2414, oct 2024.
- [16] A. Burman, J. Solé-Casals, and S. E. Lew, "Robust and memory-less median estimation for real-time spike detection," *PLOS ONE*, vol. 19, no. 11, p. e0308125, nov 2024.
- [17] W. Wang and P. Lu, "An efficient switching median filter based on local outlier factor," *IEEE Signal Processing Letters*, vol. 18, no. 10, pp. 551–554, 2011.
- [18] C. E. Graves, C. Li, X. Sheng, D. Miller, J. Ignowski, L. Kiyama, and J. P. Strachan, "In-Memory Computing with Memristor Content Addressable Memories for Pattern Matching," *Advanced Materials*, vol. 32, no. 37, 2020.
- [19] C. Li, C. E. Graves, X. Sheng, D. Miller, M. Foltin, G. Pedretti, and J. P. Strachan, "Analog content-addressable memories with memristors," *Nature Communications*, vol. 11, no. 1638, 2020.
- [20] G. Papandroulidakis, S. Agwa, A. Cirakoglu, and T. Prodromakis, "A 9t4r rram-based acam for analogue template matching at the edge," 2024. [Online]. Available: <https://arxiv.org/abs/2410.03414>
- [21] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature Electronics*, vol. 1, no. 6, pp. 333–343, 2018. [Online]. Available: <http://www.nature.com/articles/s41928-018-0092-2>
- [22] A. Sebastian, M. Le Gallo, R. Khaddam-aljameh, E. Eleftheriou, M. L. Gallo, R. Khaddam-aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnology*,

vol. 15, no. 7, pp. 529–544, 2020. [Online]. Available: <http://dx.doi.org/10.1038/s41565-020-0655-z>

[23] M. Payvand, M. V. Nair, L. K. Muller, and G. Indiveri, “A neuromorphic systems approach to in-memory computing with non-ideal memristive devices: From mitigation to exploitation,” *Faraday Discussions Royal Society of Chemistry*, 2018. [Online]. Available: <http://arxiv.org/abs/1807.05128{\%}0Ahttp://dx.doi.org/10.1039/C8FD00114F>

[24] G. Pedretti, C. E. Graves, S. Serebryakov, R. Mao, X. Sheng, M. Foltin, C. Li, and J. P. Strachan, “Tree-based machine learning performed in-memory with memristive analog CAM,” *Nature Communications*, vol. 12, no. 5806, 2021.

[25] G. Pedretti, C. E. Graves, T. Van Vaerenbergh, S. Serebryakov, M. Foltin, X. Sheng, R. Mao, C. Li, and J. P. Strachan, “Differentiable Content Addressable Memory with Memristors,” *Advanced Electronic Materials*, vol. 8-8, no. 2101198, 2022.

[26] M. Gibbs, K. Woodward, and E. Kanjo, “Combining Multiple tinyML Models for Multimodal Context-Aware Stress Recognition on Constrained Microcontrollers,” *IEEE Micro*, pp. 1–9, nov 2023.

[27] S. Stathopoulos, L. Michalas, A. Khiat, A. Serb, and T. Prodromakis, “A comprehensive technology agnostic RRAM characterisation protocol,” 2018.

[28] G. Papandroulidakis, S. Agwa, A. Cirakoglu, and T. Prodromakis, “A 9t4r rram-based acam for analogue template matching at the edge,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp. 1–14, 2025.

[29] S. Agwa, G. Papandroulidakis, and T. Prodromakis, “A 1T1R+2T Analog Content-Addressable Memory Pixel for Online Template Matching,” *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2023.

[30] A. Krizhevsky, G. Hinton *et al.*, “Learning multiple layers of features from tiny images,” 2009.

[31] L. N. Darlow, E. J. Crowley, A. Antoniou, and A. J. Storkey, “Cinic-10 is not imagenet or cifar-10,” 10 2018. [Online]. Available: <https://arxiv.org/abs/1810.03505v1>

[32] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[33] P. Kaushik, Z. Khan, A. Kajla, A. Verma, and A. Khan, “Enhancing object recognition with resnet-50: an investigation of the cifar-10 dataset,” *2024 International Conference on Smart Devices, ICSD 2024*, 2024.

[34] M. Horowitz, “1.1 Computing’s energy problem (and what we can do about it),” *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 57, pp. 10–14, 2014.

[35] J. Lin, W.-M. Chen, Y. Lin, John Cohn, C. Gan, and S. Han, “Mcunet: Tiny deep learning on iot devices,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 11711–11722, 2020.

[36] C. Banbury, V. J. Reddi, P. Torelli, J. Holleman, N. Jeffries, C. Kiraly, P. Montino, D. Kanter, S. Ahmed, D. Pau, U. Thakker, I. Antonio, T. Ii, P. Warden, J. Cordable, G. Di, G. Iii, J. Duarte, S. Gibellini, V. Parekh, V. Honson, T. V. Nhan, T. Vi, N. Wenxu, V. Xu, and X. Vii, “Mlperf tiny benchmark,” 6 2021.

[37] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized neural networks,” *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[38] V. Vasilev, V. Shterev, and M. Nenova, “Optimizing activation function for parameter reduction in cnns on cifar-10 and cinic-10,” *Applied Sciences 2025, Vol. 15, Page 4292*, vol. 15, p. 4292, 4 2025.

[39] C. He, M. Annavarapu, and S. Avestimehr, “Group knowledge transfer: Federated learning of large cnns at the edge,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 068–14 080, 2020.

[40] M. Sharif, A. Kausar, J. H. Park, and D. R. Shin, “Tiny image classification using four-block convolutional neural network,” *ICTC 2019 - 10th International Conference on ICT Convergence: ICT Convergence Leading the Autonomous Future*, pp. 1–6, 10 2019.