ASIA SOUTH PACIFIC DESIGN AUTOMATION CONFERENCE

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

先进计算机体系结构实验室 Advanced Computer Architecture Laboratory

上海期智研究院 SHANGHAI QI ZHI INSTITUTE

PAAP-HD: PIM-Assisted Approximation for Efficient Hyper-Dimensional Computing

Authors: Fangxin Liu, Haomin Li, Ning Yang, Yichi Chen, Zongwu Wang, Tao Yang, and Li Jiang*

Shanghai Jiao Tong University

饮 水 思 源 • 爱 国 荣 校

Outline

- Introduction
- Background and Related Works
- Proposed PAAP-HD
- Evaluation
- Conclusion

Introductions: HDC Limitations

TABLE I
PREFORMANCE OF HDC WITH DIFFERENT KINDS OF DATA BIT WIDTH.

Metrics	Accuracy(%)			Energy Consumption(μJ)	
Configuration	FP32	INT8	Binary	INT8	Binary
ISOLET [6]	91.53	90.89	67.73	8.46	2.82
HAR [7]	90.83	89.97	59.14	4.445	2.407
CARDIO [8]	84.98	83.10	68.08	0.558	0.186

[a]Evaluation is performed on FPGA with Verilog implementation.

Binary HDC:

Data precision ↓

Energy efficiency ↑

high-precision cosine distance calculations:
Large Computational Cost

? HDC in edge computing

Introductions: Our Contributions——PAAP-HD

PIM-assisted Approximate Model

- An effective alternative model for HDC
- Map the approximate model to parallel crossbar-based MAC operation

Predictor Design for Module Switching

- Enabling a seamless transition between the HDC and approximate models

Comprehensive Evaluation

- Across various datasets from different fields

HDC Basics

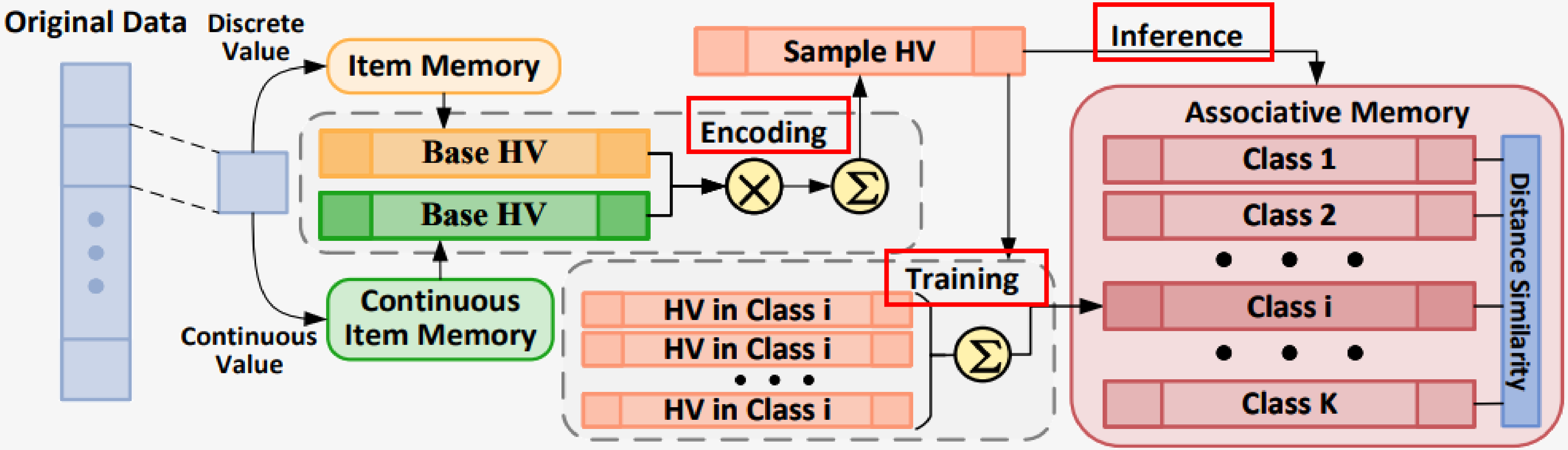


Fig. 1. HDC classification framework. HV is short for hypervector.

- Three Basic Modules: IM ,CIM,AM
- Three Stages for Learning: Encoding, Training, Inference
- Binary HDC: binarize each class hypervector in AM &the sample hypervector

Computing Basics

Approximate Model

- Introduced a simple predictor

In-memory Computing

- ReRAM-based PIM

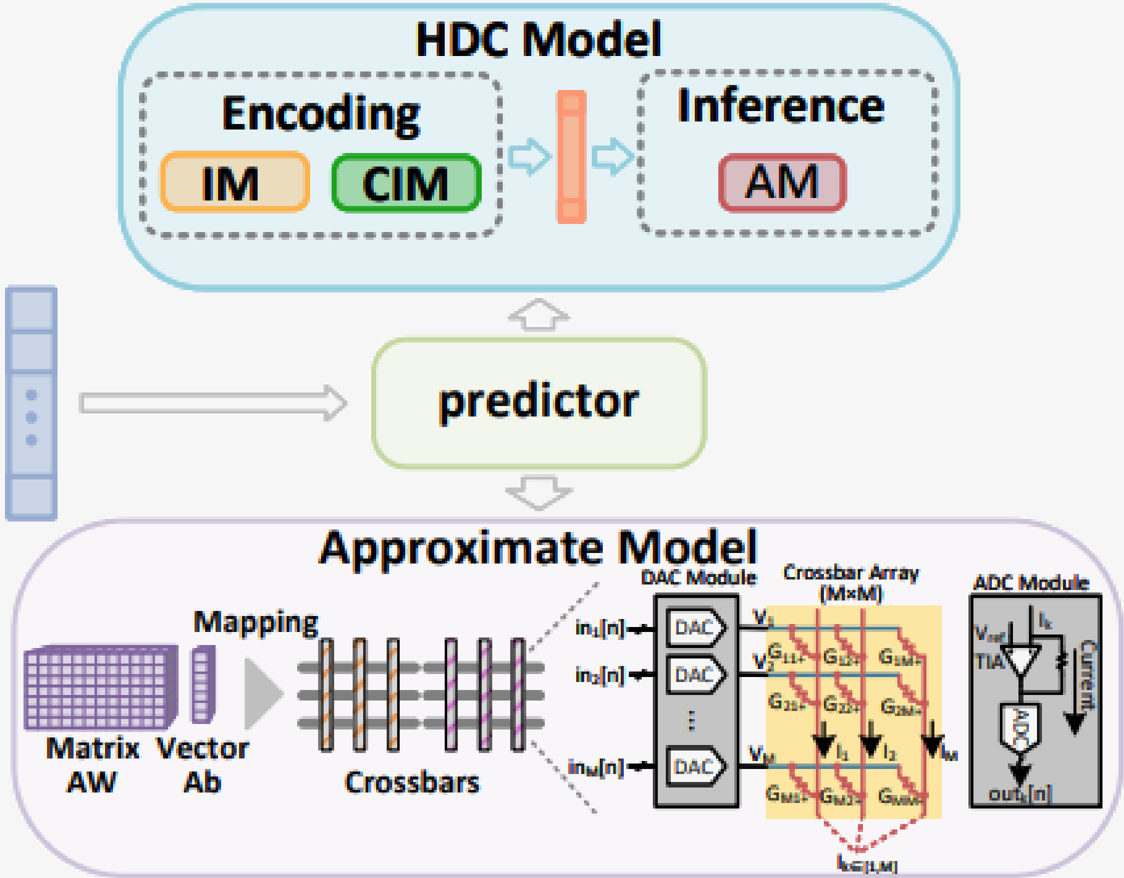
Motivation

- Not fully leveraged the direct benefits of PIM
- Tailored solely for binary HDC

Overview of Proposed PAAP-HD

3 Main Components:

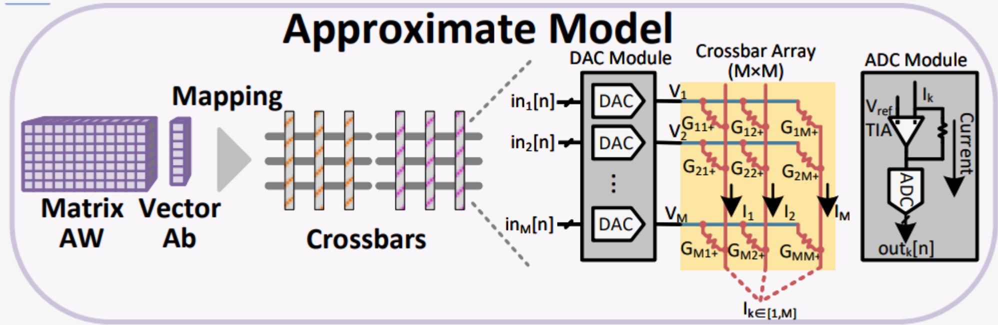
- A high-precision HDC model
  - **INT8** HDC model
- An approximate model
  - **mapped** onto a PIM-based crossbar
- A quality control predictor
  - **determine** which model to use



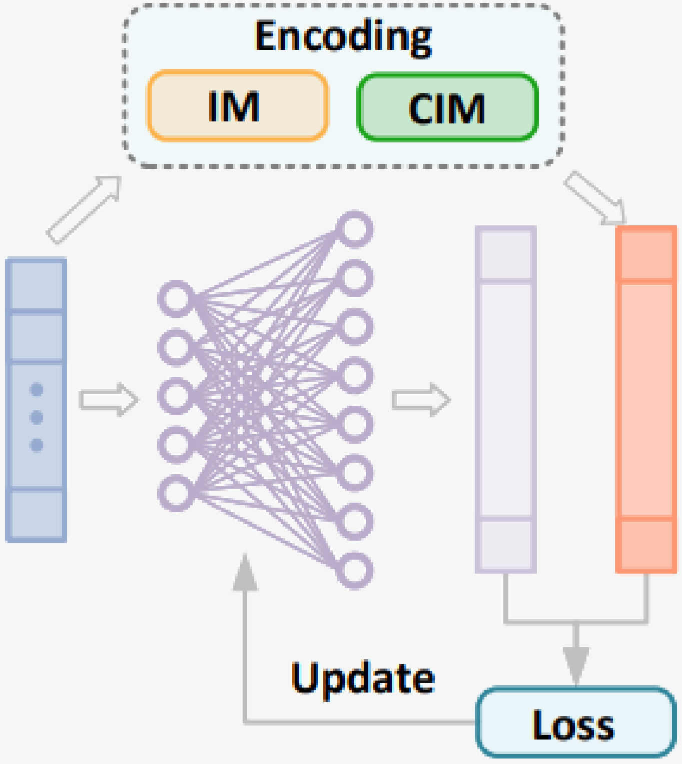
(a) Overview of HyperApprox

Approximate Model Design

- The associative search in the AM
  - $prediction = \underset{i=1,\ldots,K}{arg\,max}\, Cos(\vec{S},\ \vec{C_i}) = \frac{\vec{S}\cdot\vec{C_i}}{|\vec{S}|\cdot|\vec{C_i}|}$
- The neural approximation
  - $Dotp\ \ result = A(W\vec{f} + \vec{b}) = AW\vec{f} + A\vec{b}$
- Trained on the output

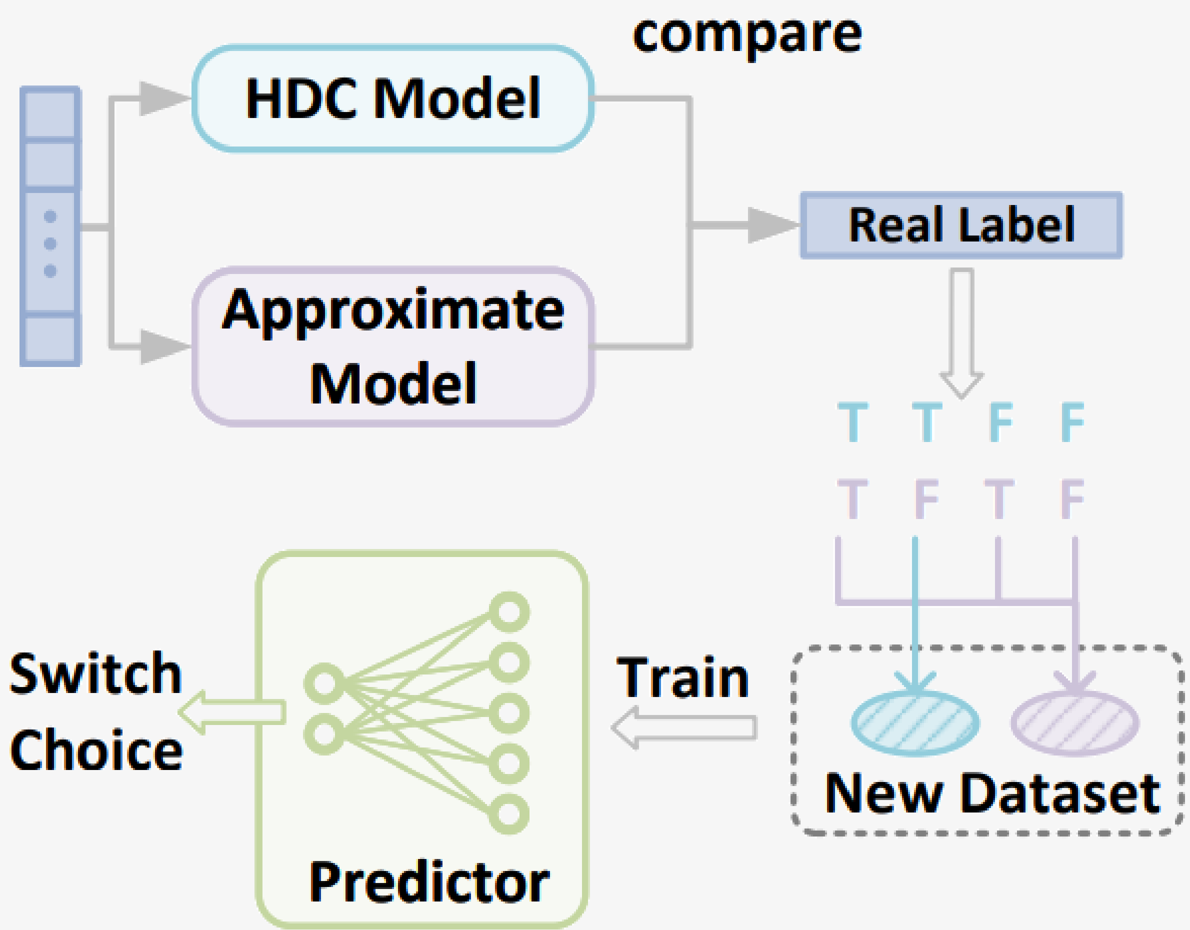


Fully leverage the benefits of PIM



(b) Approximate Encoding

Design of a Quality Control Predictor



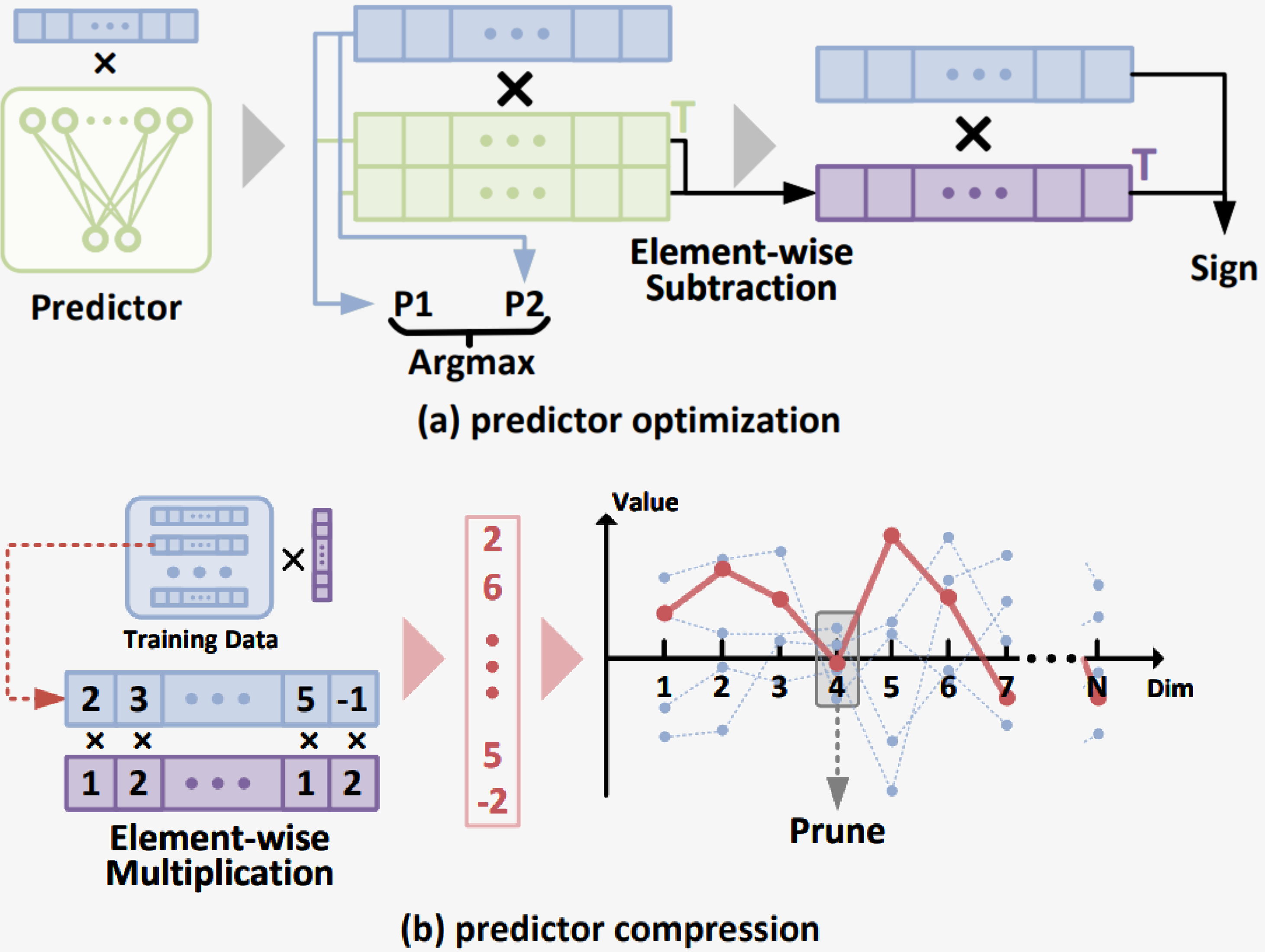
(c) Predictor Design

the Predictor

PAAP-HD

a MLP
without hidden layers and biases

Optimization and Compression of Predictor



Evaluation: Experimental Setup

Implementations

- Software: utilized Python on an Intel(R) Xeon(R) Silver 4208 CPU with 16GB memory
- Hardware: involved a Kintex-7 FPGA and Verilog for predictors and high-precision HDCs

Datasets and Models

- Speech recognition, activity recognition, and disease classification
- Main comparison: PAAP-HD and the traditional HDC model (models set: 1, 000)

Metrics

- Acc
- Extr … accuracy

TABLE II
DATASET STATISTICS

	#feature	#class	Train size	Test size	Description
ISOLET [6]	617	26	6,238	1,559	Speech recognition
UCIHAR [7]	561	12	6,213	1,554	Activity recognition(Mobile)
CARDIO [8]	21	10	1,913	213	Disease classification

Evaluation: Accuracy

TABLE III
ACCURACY COMPARISON

Configuration	Accuracy		
	binary HDC	INT8 HDC	PAAP-HD
ISOLET	67.73%	90.89%	90.38%
HAR	59.14%	89.97%	89.34%
CARDIO	68.08%	83.10%	81.69%

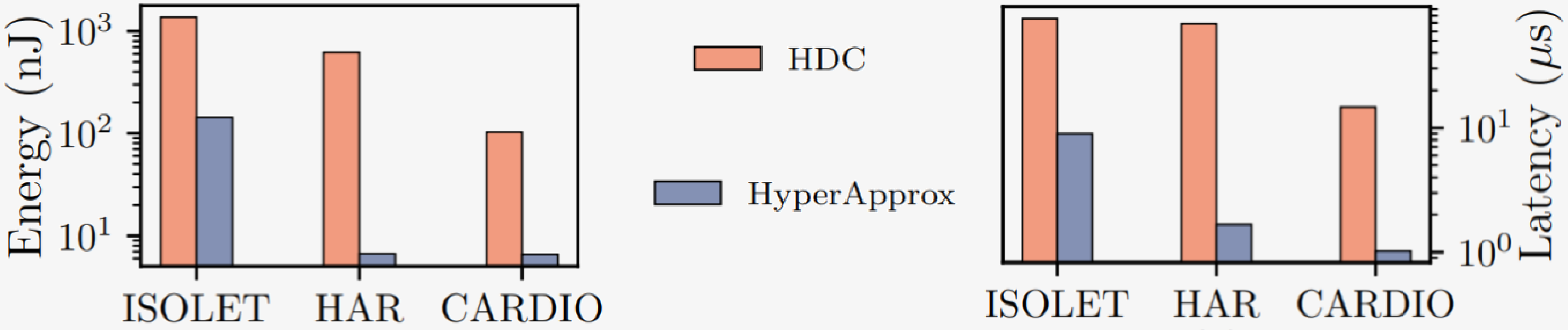
Employed the INT8 HDC version as the baseline

Compared to the original HDC model:

Speech and action recognition tasks: < 0.5%

Disease classification task: < 1%

Evaluation: Overall Performance



(a) (b)

Energy Efficiency

An energy efficiency improvement of 93.02×

An average energy efficiency gain of nearly 39.42×.

Performance

A speedup of 21.37× on average

reduced computation
low latency

Evaluation: Approximation Details

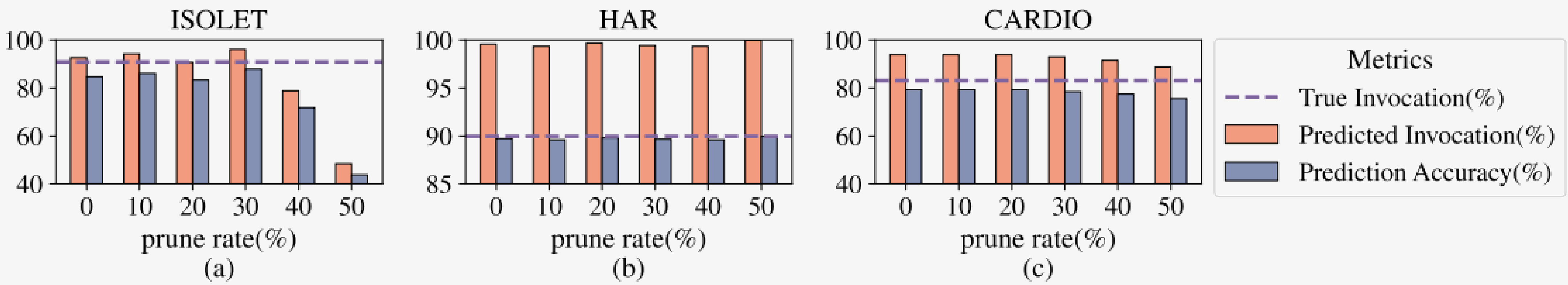


Fig. 5. Predictor's performance on three datasets with different dimension-pruning rate.

TABLE III
ACCURACY COMPARISON

Configuration	Accuracy		
	binary HDC	INT8 HDC	PAAP-HD
ISOLET	67.73%	90.89%	90.38%
HAR	59.14%	89.97%	89.34%
CARDIO	68.08%	83.10%	81.69%

Nearly all samples deemed approximable for the HAR dataset

NO negatively impact

Evaluation: Approximation Details

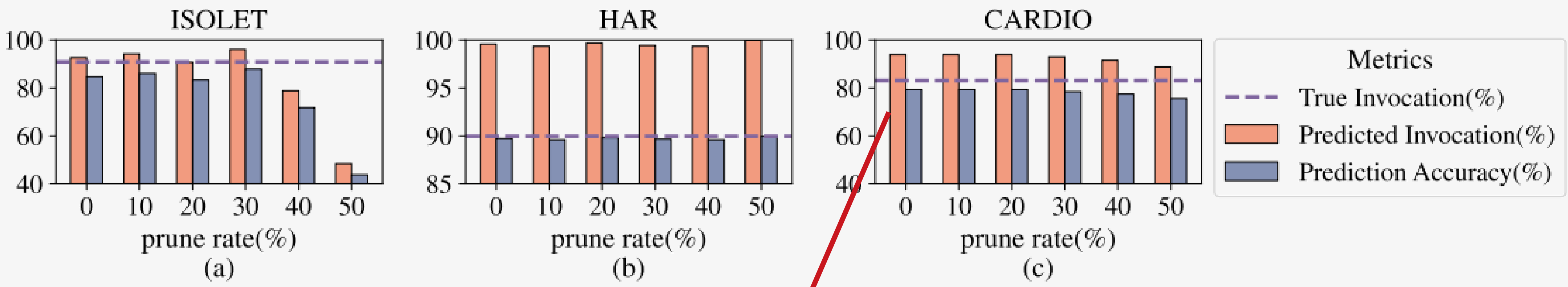


Fig. 5. Predictor's performance on three datasets with different dimension-pruning rate.

A slightly lower final accuracy than the baseline

On the HAR and CARDIO datasets:
Extensive pruning of predictor dimensions —— minimal impact

On the ISOLET dataset:
Pruning up to 30% of dimensions——slightly influences

Evaluation: Variation Analysis

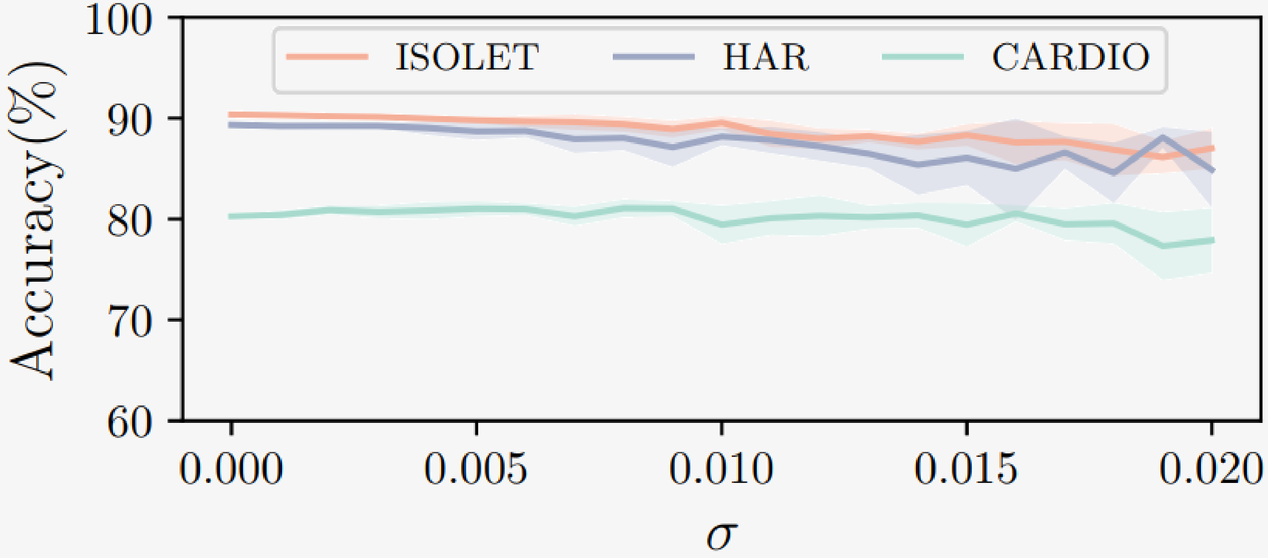


Fig. 6. Variation performance with different σ.

Challenge: Resistance Variations

σ = 0.02: no accuracy loss

All σ values: consistent level of accuracy
(Speech recognition & disease classification tasks)

Conclusion

- PAAP-HD: a PIM-assisted framework for energy-efficient hyperdimensional computation
  - Fuses and approximates the encoding and association search stages of HDC
  - Integrating specially designed approximate models and quality control predictors
  - Minimize energy consumption
- Enhances HDC inference efficiency
  - the potential of approximation
  - Significant advancement in energy-efficient HDC design.

Thank you!

PAAP-HD: PIM-Assisted Approximation for Efficient Hyper-Dimensional Computing

Authors: Fangxin Liu, Haomin Li, Ning Yang, Yichi Chen, Zongwu Wang, Tao Yang, and Li Jiang*

Shanghai Jiao Tong University