Robust and Practical Machine Learning Solutions for Wearable Biomedical Edge Devices

Integrated Systems Laboratory (ETH Zürich)

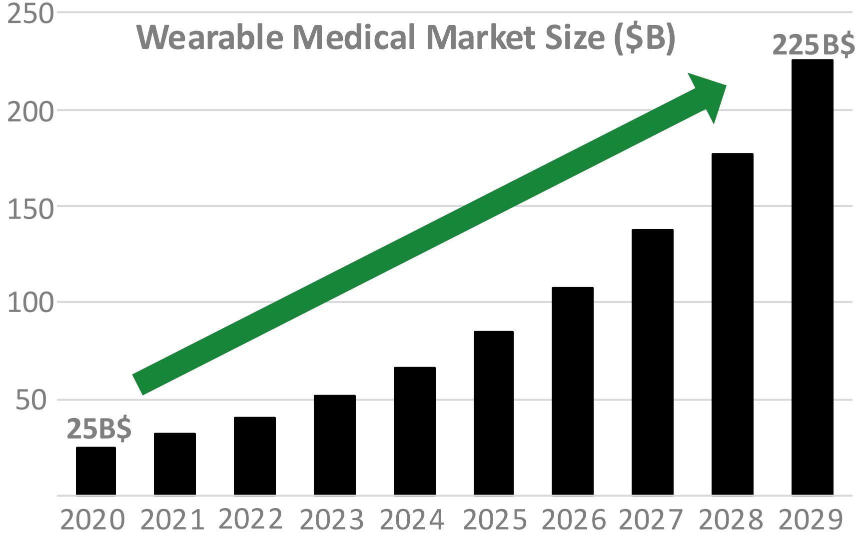
Thorir Mar Ingolfsson

PhD Examination (Diss.-No. 31324), July 1, 2025

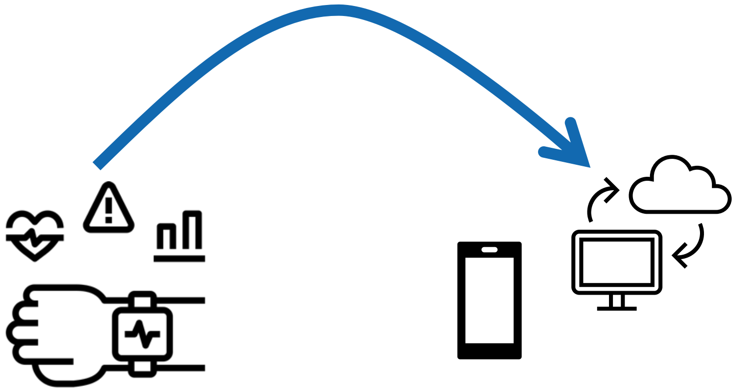
Chair: Prof. Dr. Janos Vörös
Examiner: Prof. Dr. Luca Benini
Co-examiner: Prof. Dr. Mauro Mangia

The Wearable Revolution & Its Processing Gap

Rapid Growth in Wearable Health Monitoring

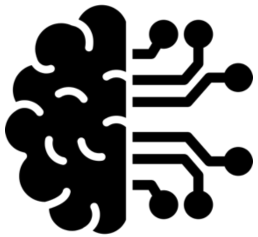


Heart Rate (**ECG,PPG**), Brain Signals (**EEG**) etc.

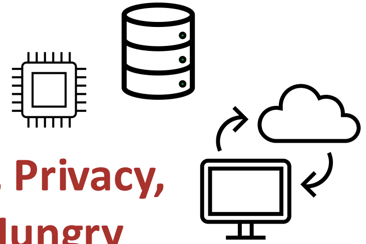


Wearables **gather data** and **send away**

Edge devices



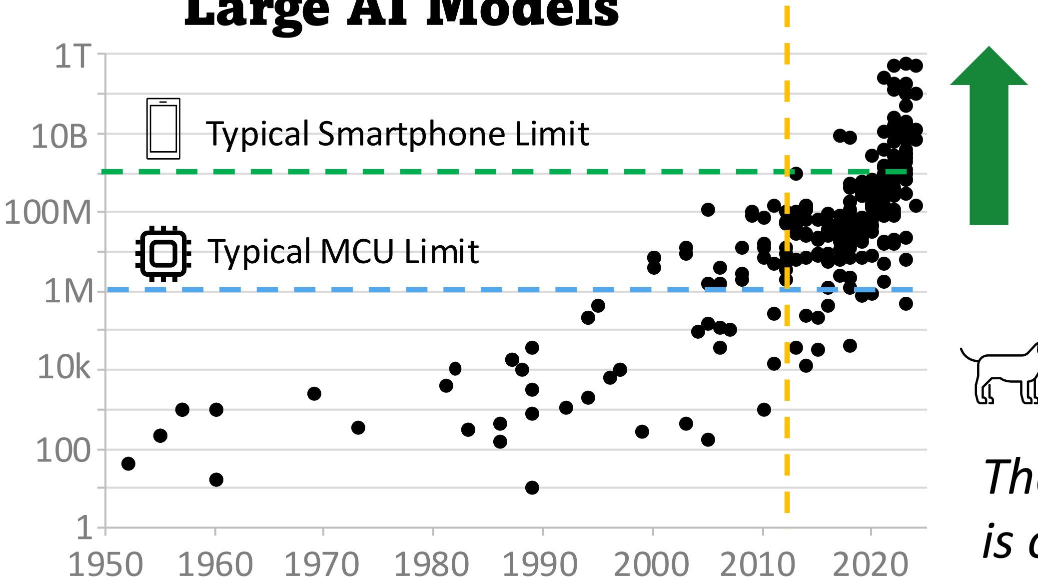
Limited On-device processing!



Latency, Privacy, Power Hungry

A Widening Chasm: Powerful AI vs Tiny Devices

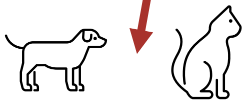
Increasingly Powerful & Large AI Models



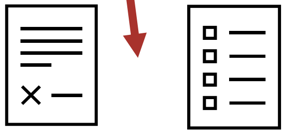
2012
60M

30,000x

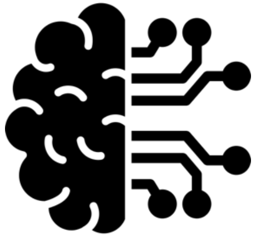
2024
1.8T



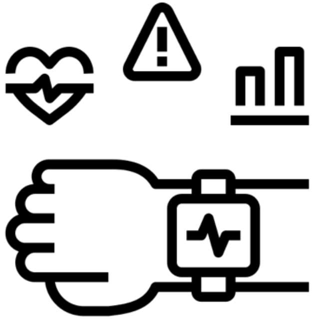
That image is of a cat



Here is your entire medical history analyzed



How do we push these **large brains** onto **small battery** powered **edge devices?**



Bridging the Gap with Practical, Robust and Scalable AI

Resource efficiency

Strict **computational budget**.

Practical

Days/weeks runtime

Signal quality and noise

Real world bio-signals are **messy**.

Robust

Tackle noise head on

Adapting to a changing world

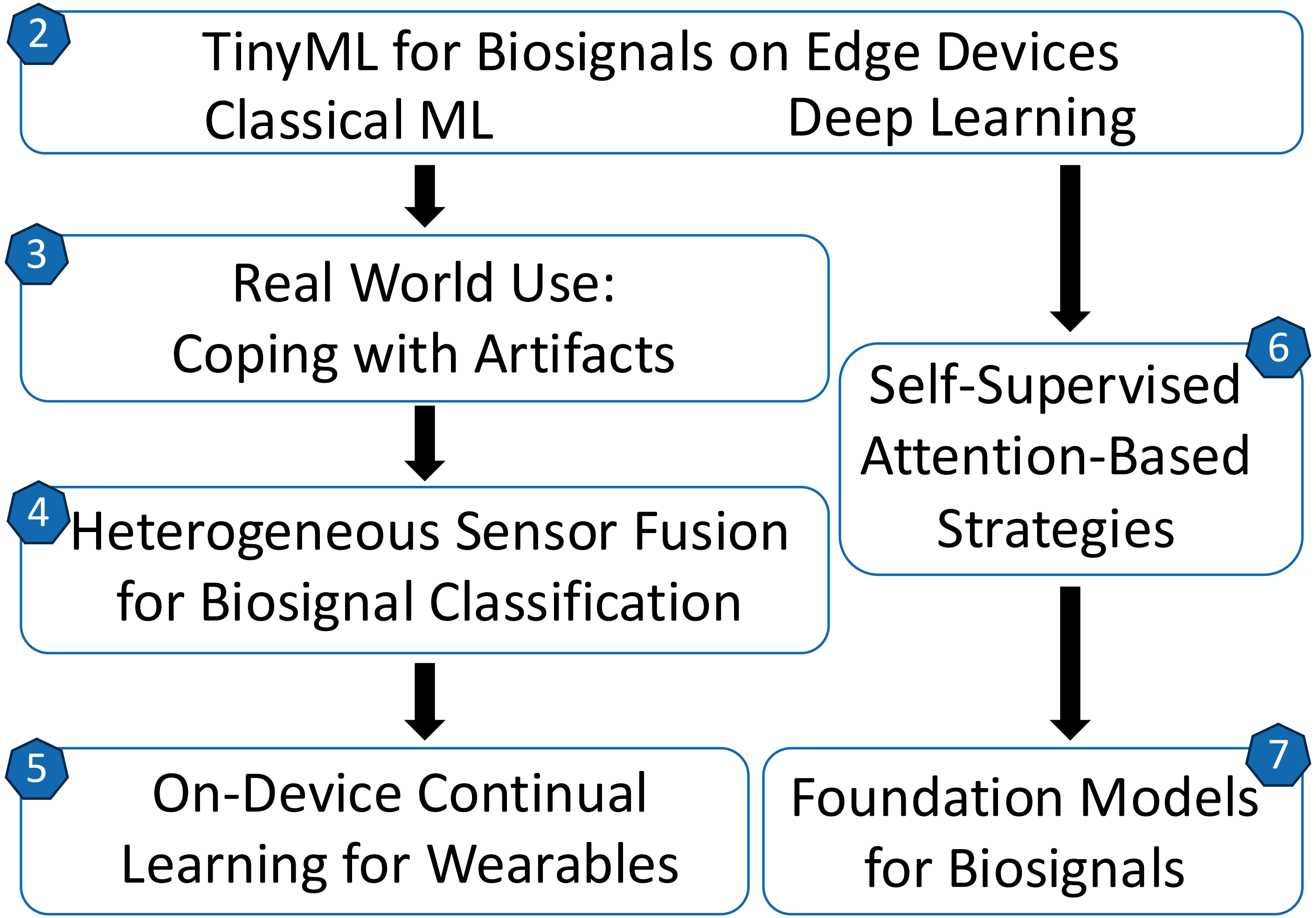
Signals **evolve** and **change** over time

Scalable

Continual adaptation & multi-domain use

And to do these things we need **HW <-> SW Co-design**

The Resarch Roadmap: A Path to Universal Models



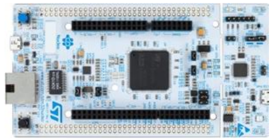
Algorithms for HW and HW for Algorithms

Multiple platforms explored & optimized

Apollo4

GAP9

GAP8



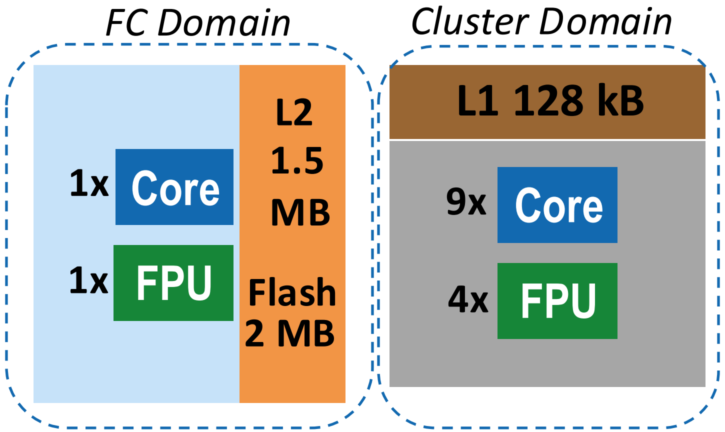
STM32F756

STM32L475

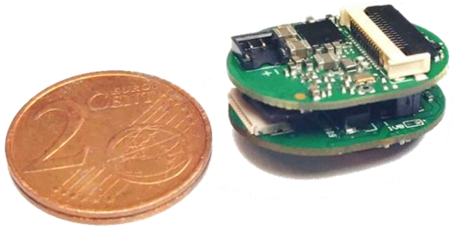
All explored

Most energy efficient

To stay ahead of the curve



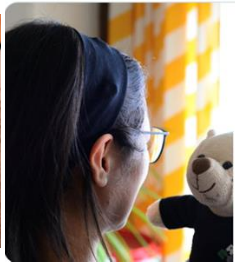
BioGAP

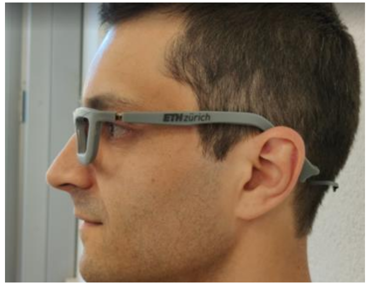


Key limitations:

Memory: **L1 128kB**
L2 1.5MB

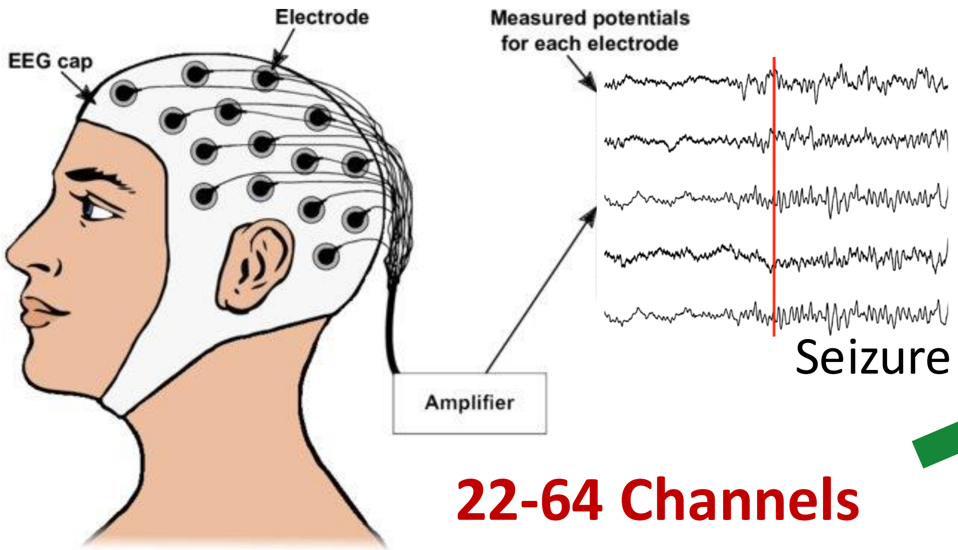
Speed: **400MHz**
9 Core computation cluster





Seizure Detection Use-Case: EEG as a Challenging Biosignal

- Electroencephalography **(EEG)**: 10 µV to 100 µV ← low SNR



22-64 Channels

We have less channels

Wearable solutions
4 Channels

Temporal region

Sensitivity: Seizure Windows Detected ↑

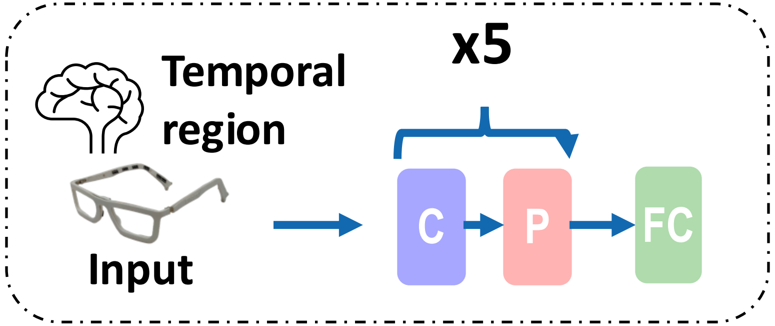
Specificity: Normal Windows Detected ↑

False Positives (FP): Wrong Alert of a Seizure ↓

Schweizerischer Nationalfonds

PEDESITE

EpiDeNet: An Efficient Foundation for On-Device Seizure Detection



✓ Operates on ~~23~~ only **4 channels**

✓ Hardware and algorithmic co-design

✓ Compact and Efficient, 11K parameters

EpiDeNet

Sens. ↑

Spec. ↑

FP. ↓

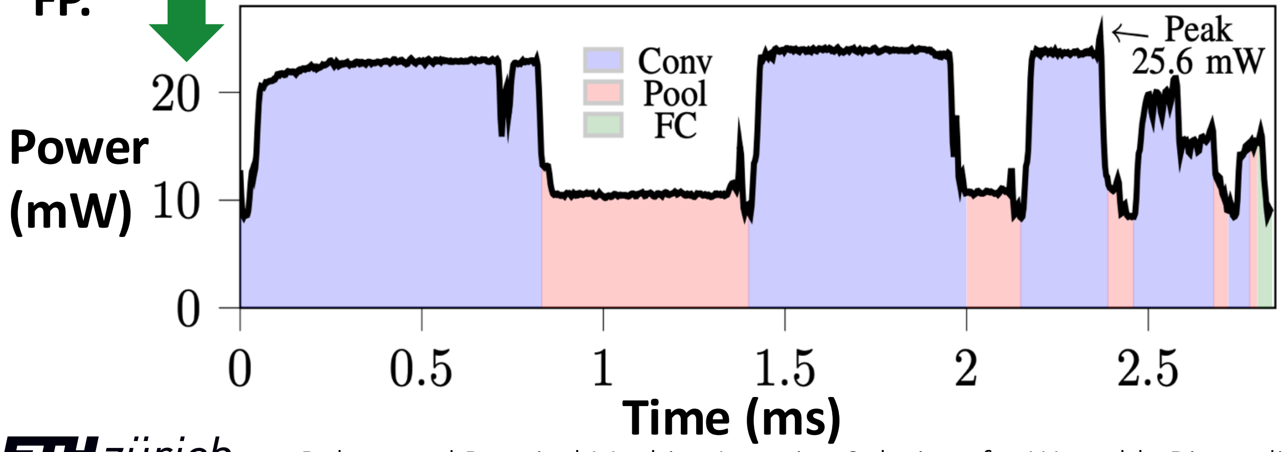
Sensitivity **S**pecificity **W**eighted **C**ross **E**ntropy Loss (**SSWCE**)

$$\mathbf{SSWCE}(\mathrm{y},\mathrm{p}) = \mathbf{CE}(\mathrm{y},\mathrm{p}) + \alpha(1-\mathbf{SP}) + \beta(1-\mathbf{SN})$$

94% Seizures detected

2.28 FP/h

Further reduced to **0.51 FP/day** using Sensor fusion



0.051 mJ per inference
300mah battery

→ **300 hours**

Work	Approach	Time	En. Eff. [GMAC/s/W]
[Liu et al., 2021]	LSTM	100ms	16.00
[Busia et al., 2022]	Transformer	405ms	3.51
Us	**CNN**	**2.84ms**	**40.61**

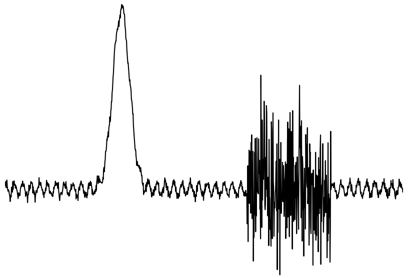
The Real-World Hurdle: The Dominance of Artifacts

#1

SoA models are trained with **clean**, clinical grade data

Very high Sens, Spec, and low FP.

✓ Data quality (SNR) **high, no artifacts**



✗ Data quality (SNR) **low, artifacts**

False alarms

Wearable, real-world data is messy

False alarms predicted by models

Predicting a seizure when there is no seizure present.

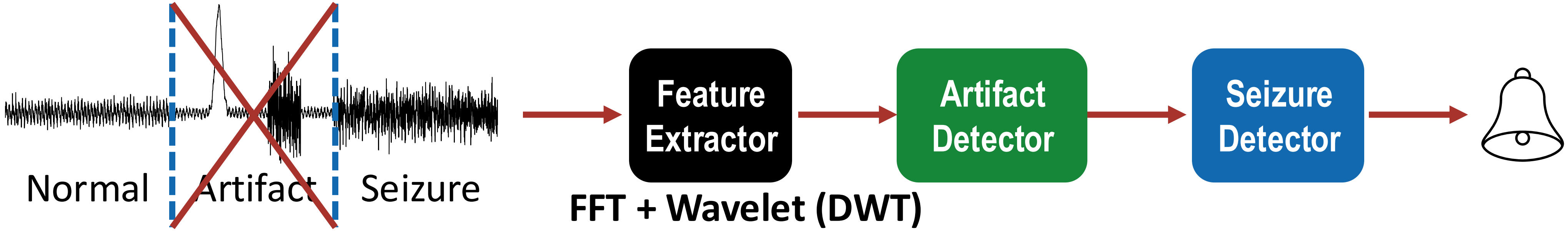
> 1 FP/hour

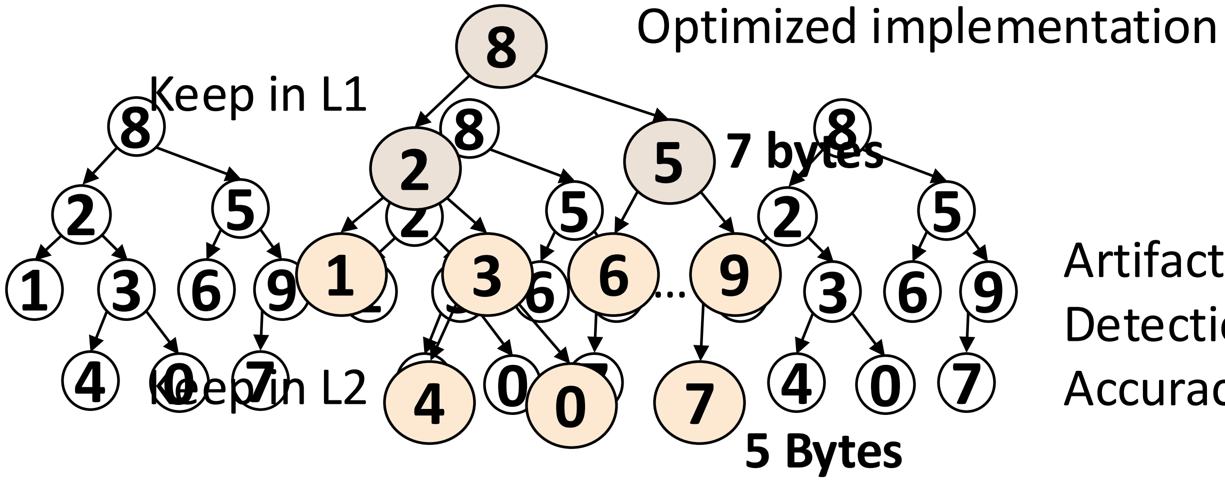
< 1 FP/Day

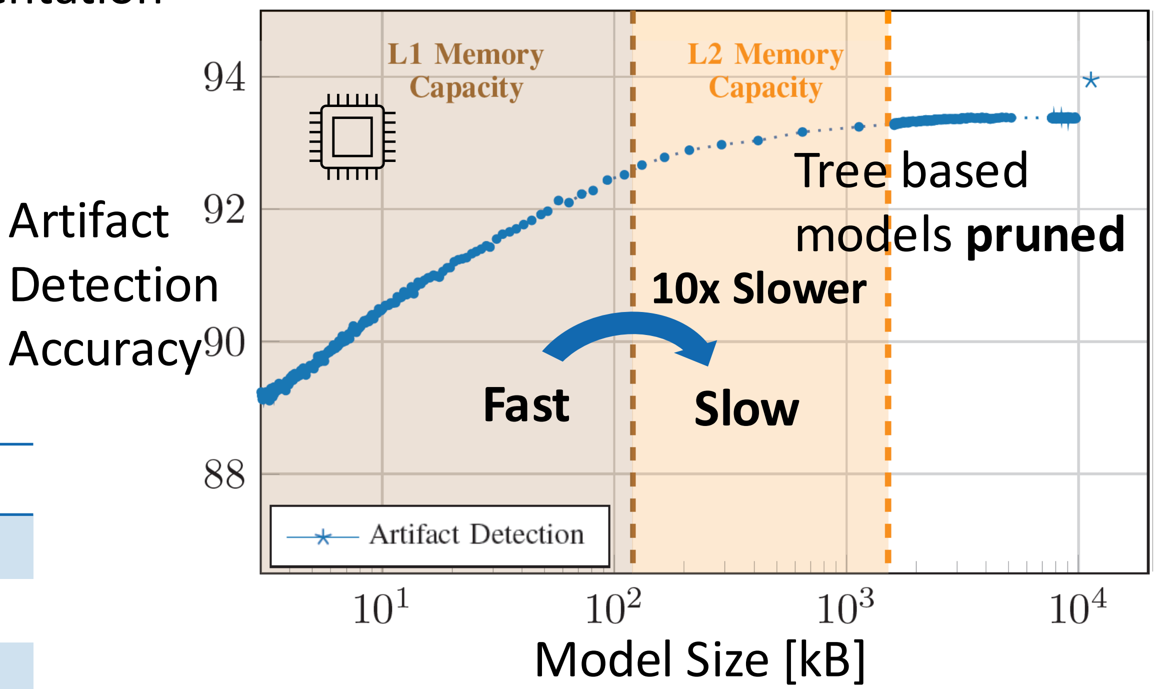
Schweizerischer Nationalfonds

PEDESITE

Fitting Decision Trees On Limited Memory While Reducing False Alarms by 96%

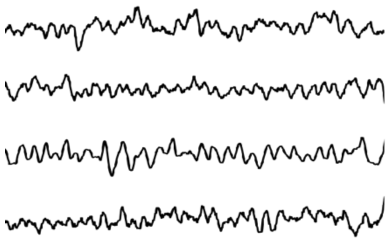






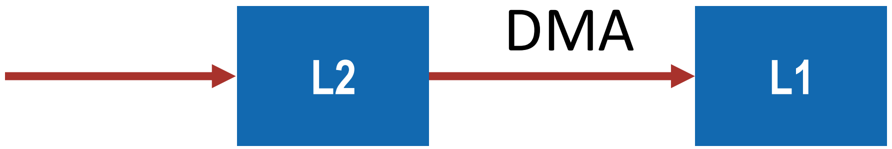
Work	Approach	False Alarm Reduction
[Islam et al., 2016]	Cleaning	80%
[Islam et al., 2020]	Cleaning	49%
Us	**Detection**	**96%**

Hardware validation, this is practical not just theoretical.



Data streaming in at 250-500 Hz

16/32 kB per 4s window (4 channels)



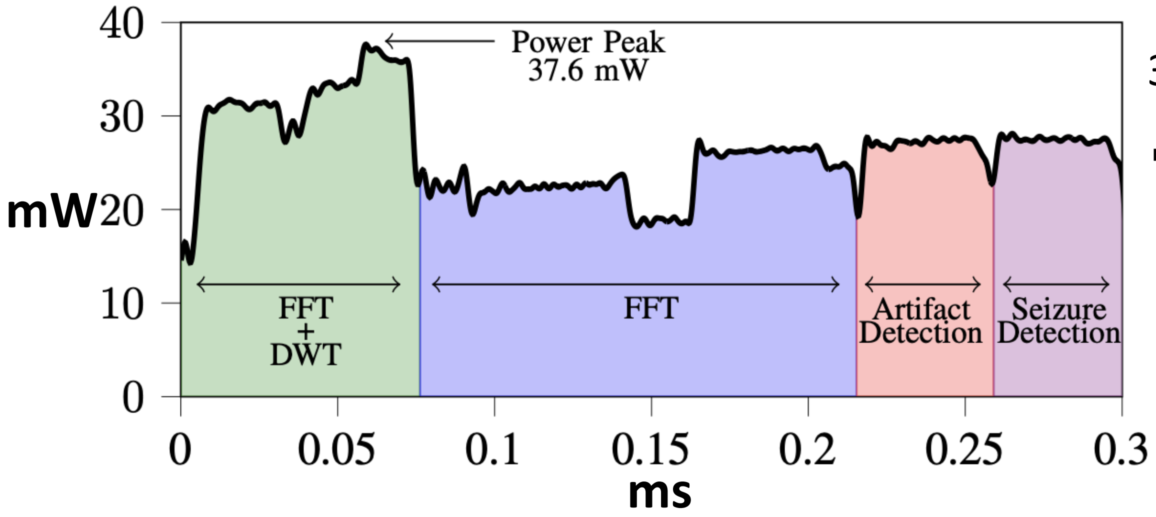
Feature extraction

- 4 Cores FFT Features
- 4 Cores DWT Features
- 1 Core for memory transfers

Artifact/Seizure Detection

# of trees Divisible by number of Cores

7.93 µJ energy per 4 s window



300mah battery

→ **12.5 days**

GAP9 Implementation

Artifact Detection in Live-Demos for BCI Drone Control

The Next Challenge: Adapting to Evolving Biosignals

But Biosignals **change with time**

Newly trained model

Good performance

Same model, new data

Bad performance

Fast, **Expensive**, **Privacy**, **Latency**

Slow, **Privacy**, **Latency**

Do the training **On-device**

Buffers

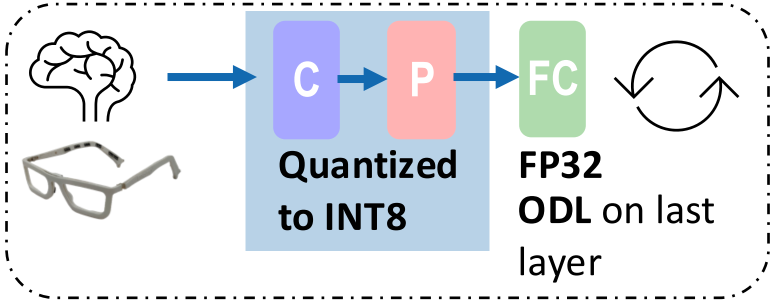
Code Data

Limited space

L1 128 kB

...

L2 1536 kB



18.6 kB of additional storage
One training step in **22 ms** at a **15mW** power budget

Minimal effect on battery life

Work	#Params.	Meas.	Accuracy
[Ma et al., 2022]	291.6k	✗	79%
[Lee et al., 2023]	8.9k	✗	71%
Us	**7.7k**	✓	**91%**

30% increase in accuracy (BCI use-case)

The Next Frontier: From Specific to Universal Models

X: Raw Data
Cheap

Heart Rate (ECG)
Brain Signals (EEG)
Muscle Activations (EMG)

~ 10TBs (20k+ hours).

Y: Labels
Expensive

Atrial Fibrillation, Seizure, etc.

Supervised Learning

X, Y → **Training** → **Model**

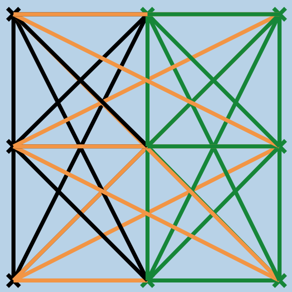
Self-Supervised Learning

X → **Training** → **Model** →

Foundation

A Toolkit for Efficient Foundation Models

CEREBRO



Full Attention

$O(N^2)$

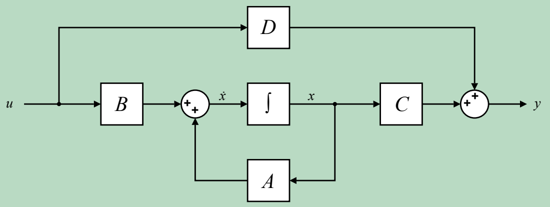
Spatial

Alternating Attention

Temporal

Solves Spatio-Temporal Complexity

FEMBA



Utilizing State Space Models

$O(N^2)$ → $O(N)$

Linear Scaling

LUNA

8-128

Widely Varying #Channels

C → Channel Unification Module → Q

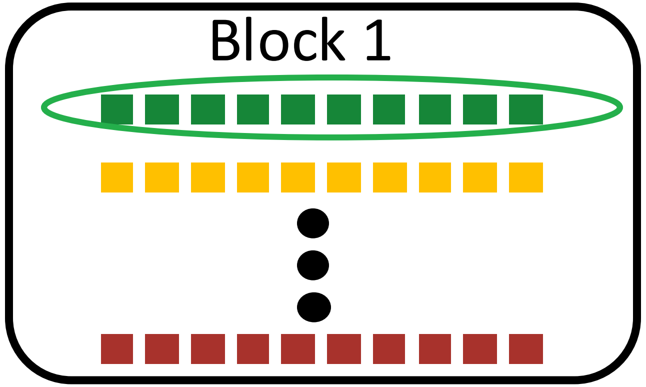
Channel Invariance

Innovation 1: CEReBrO's Alternating Attention

Standard Self Attention → **Expensive**

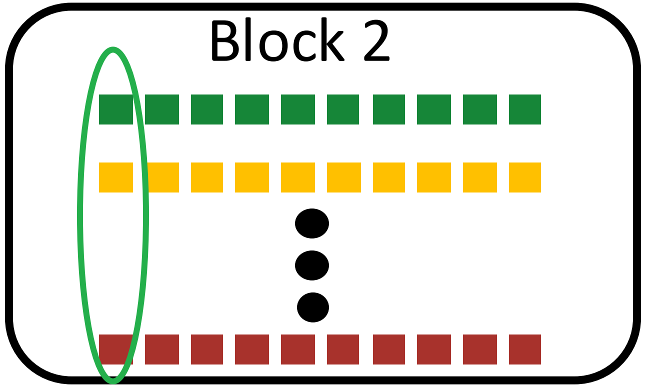
$O(N^2)$

$O(TN + CN)$



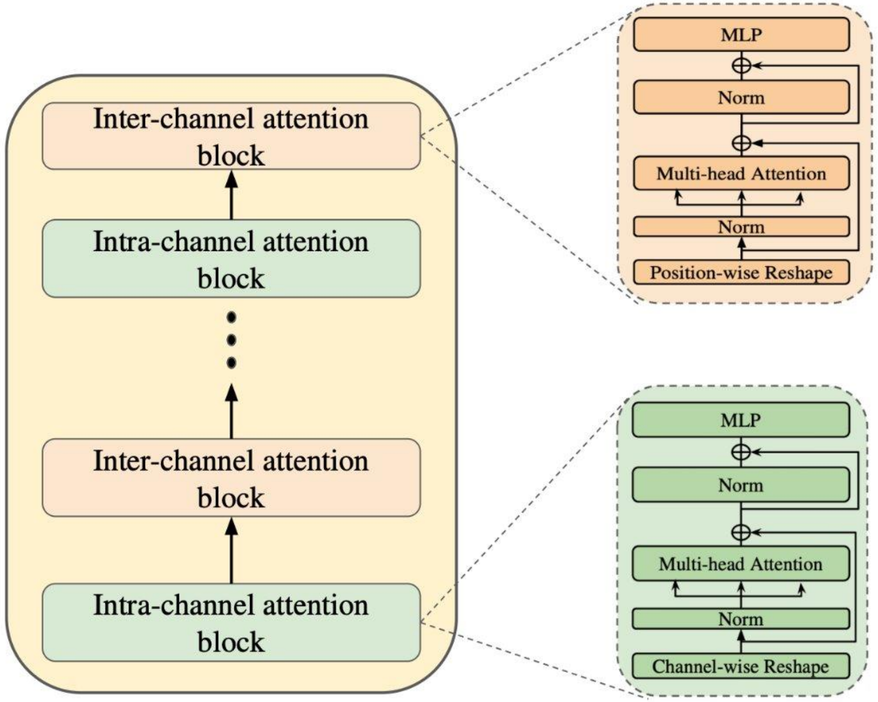
Intra-Channel Attention

Self-attention between tokens of the same channel



Inter-Channel Attention

Self-attention between tokens from the same timestep



Innovation 2: State Space Models for Linear Scaling

Attention is still quadratic in nature

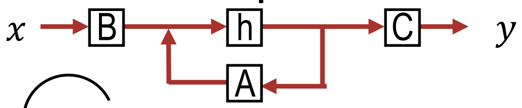
$$O\left(\frac{N^2}{C} + CN\right)$$

State: $h_k = Ah_{k-1} + Bx_k$

Output: $y_k = Ch_k$

Recurrence

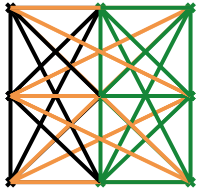
State Space Models



O(N)

Kernel Trick + Parallelizable operations → Structured State Space Models → **Mamba**

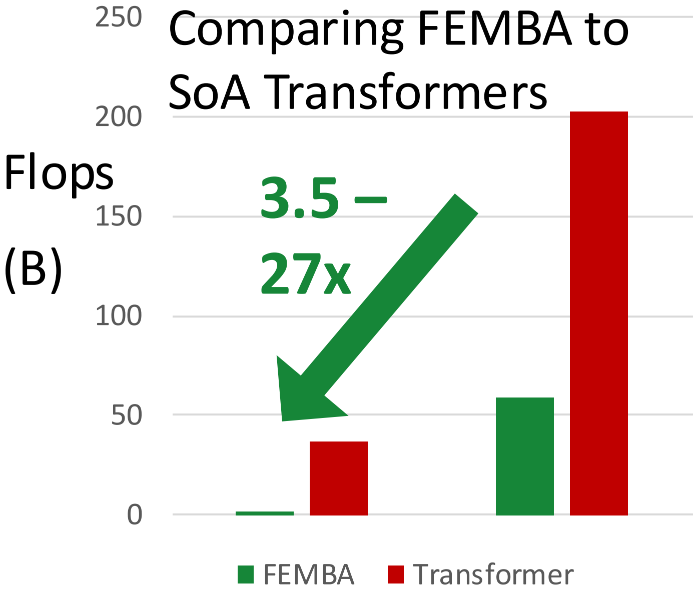
Transformers:

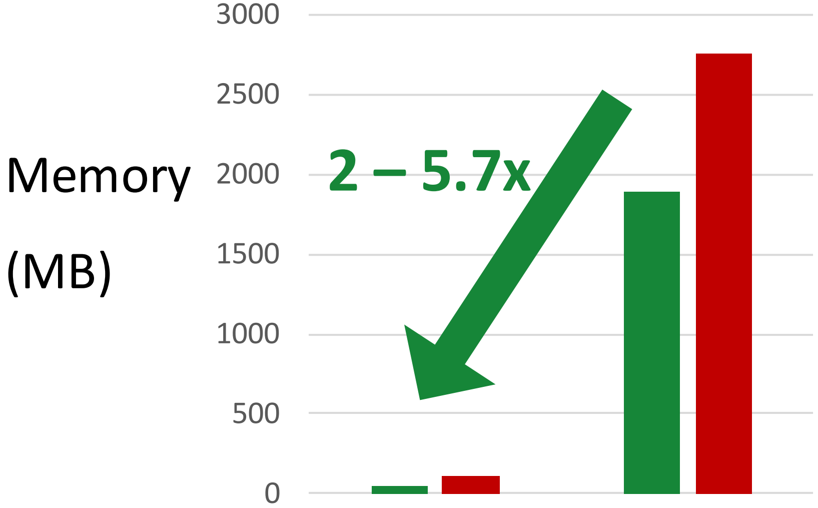


O(N²)

NxN Matrix

Comparing FEMBA to SoA Transformers

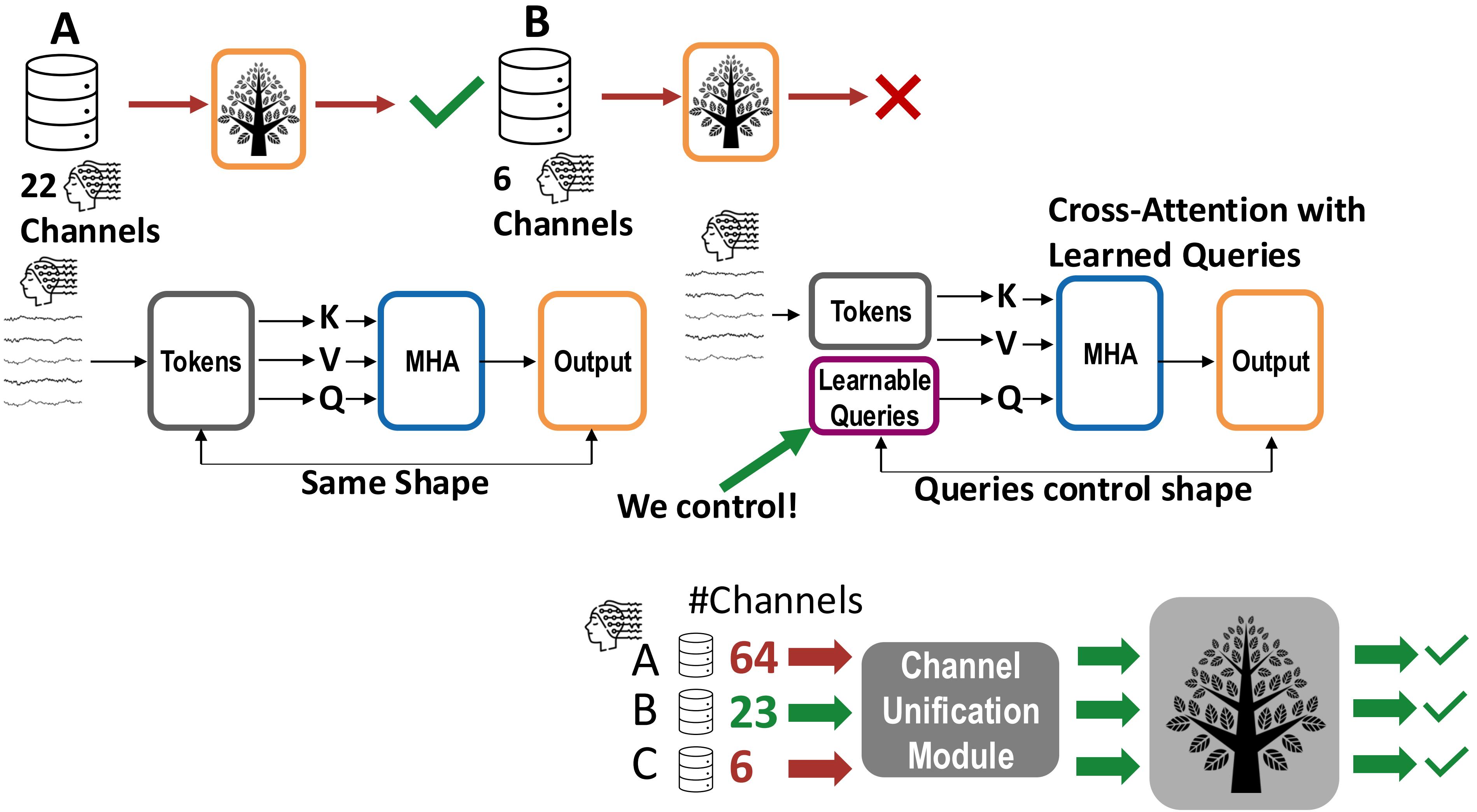




Fading Memory

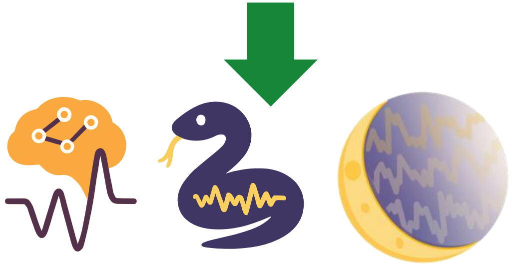
Not as optimized

Innovation 3: Learned Queries for Topology-Invariance



SoA Performance at a Fraction of the Cost

Terabytes unlabeled data



Alt-Attn, Mamba, Channel Unification

SoA in 4 downstream tasks

Model	Param. (M)	FLOPs (B)		AUROC	
[Jiang et al., 2024]	46	28		0.913	SoA Transformer based Model
CEReBrO	**40**	**18**	**↓1.5x**	0.887	**Abnormal EEG Detection**
FEMBA	**47**	**7**	**↓4.0x**	0.883	
LUNA	**43**	**8**	**↓3.5x**	0.883	

Close to SoA results at a cheaper price!

Model	Param. (M)	FLOPs (B)		AUROC	
[Chen et al., 2024]	3.3	36.4		0.852	SoA Transformer based Model
FEMBA	**7.8**	**1.3**	**↓28x**	**0.918**	**EEG-Artifact Detection**

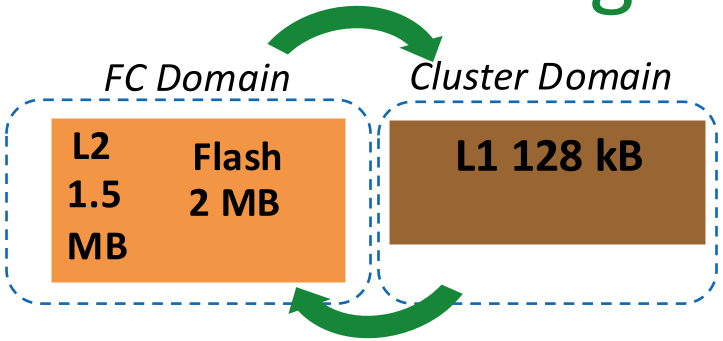
Bringing Foundation Models to the Edge

CEReBrO

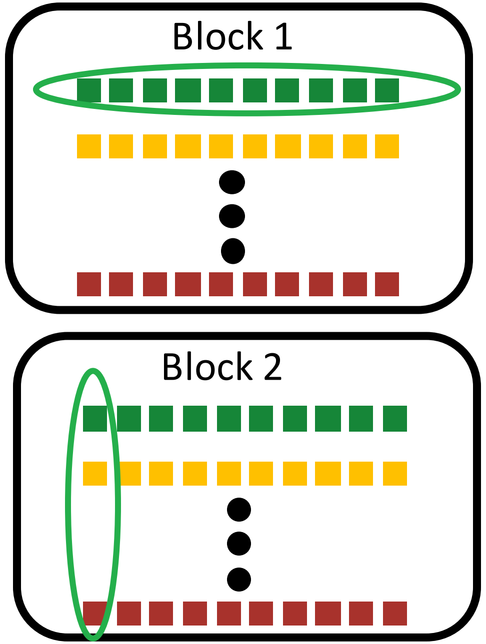
Smallest model:
2.5M parameters

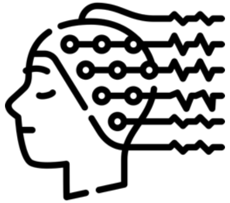
GAP9

9.5 MB in FP32 → **2.3 MB in INT8**



L3 → L2 → L1
Tiling





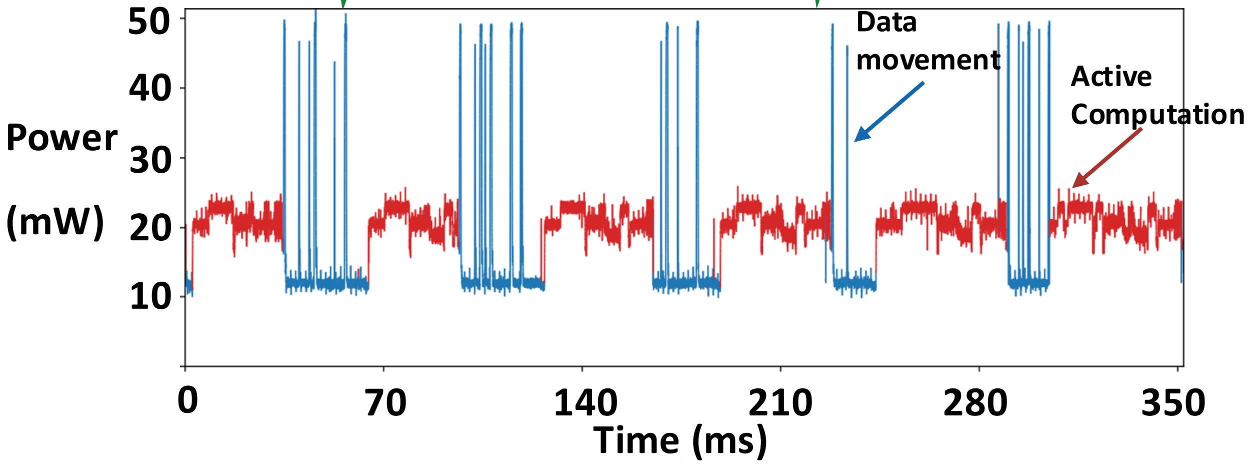
16 EEG channels

2 seconds of data

Specific kernels written to handle alternating attention

10 mJ per inference

0.35s Runtime!



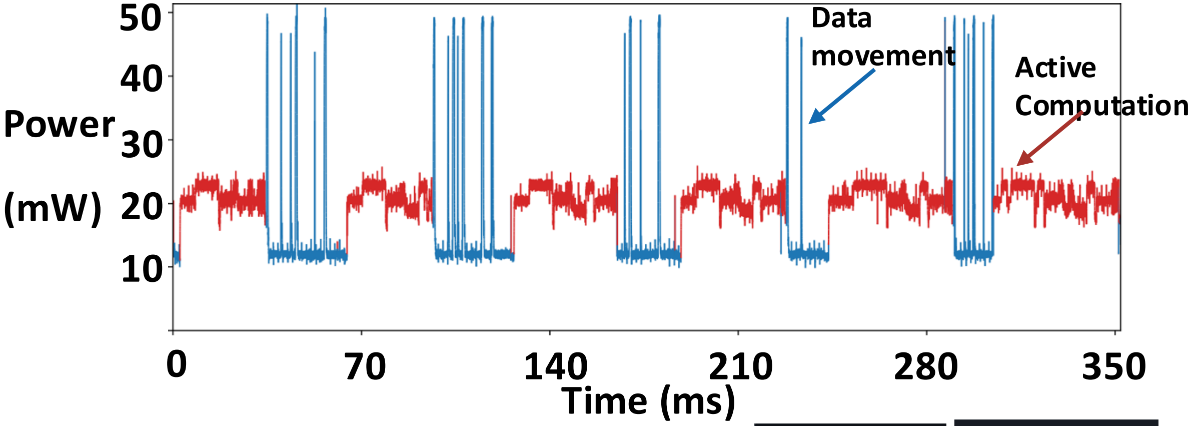
Summary of Key Contributions

Developed Robust & Practical ML Pipelines for Wearable Biosignal Analysis

- Classical ML — **Low-channel/Subject Specific**
- Artifact Mitigation — **96%** ↓ **of false alarms**
- Sensor Fusion — **<1FP/Day**
- Domain-Specific Optimization — **SSWCE**

Designed and Deployed Ultra-Low-Power Models for On-Device Inference

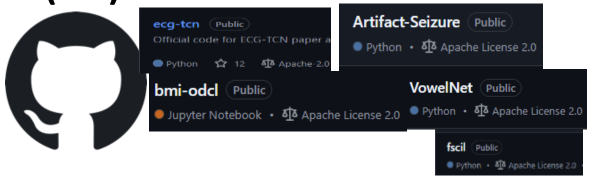
- Energy-Efficient architectures — EEG-TCN, EEGFormer, EpiDeNet, **BrainFuseNet**
- Hardware Validation — **<0.11 mJ** on GAP9
- On-Device Continual Learning — **30%** ↑ at 0.5 mJ/update



Pioneered Efficient Foundation Models for General-Purpose Biosignal Representation

- Novel Architectures Introduced
- Cross-Modal Transfer — **EEG → ECG/PPG**
- First time Deployment on Edge Device

Open-Sourced to the Community



Big Thanks to All Collaborators

[1] **T. M. Ingolfsson**, V. Kartsch, L. Benini and A. Cossettini, "A Wearable Ultra-Low-Power System for EEG-based Speech-Imagery Interfaces," in IEEE Transactions on Biomedical Circuits and Systems, 2025

[2] B. Döner**, T. M. Ingolfsson**, L. Benini, Y. Li, "LUNA: Efficient and Topology-Agnostic Foundation Model for EEG Signal Analysis", 1st ICML Workshop on Foundation Models for Structured Data, 2025

[3] B. Tóth, D. Senti **,T. M. Ingolfsson**, et al., "Finetuning and Quantization of EEG-Based Foundational BioSignal Models on ECG and PPG Data for Blood Pressure Estimation ", IEEE Engineering in Medicine & Biology Society (EMBC), 2025

[4] A. Tegon, **T. M. Ingolfsson**, et al.,"FEMBA: Efficient and Scalable EEG Analysis with a Bidirectional Mamba Foundation Model", IEEE Engineering in Medicine & Biology Society (EMBC), 2025

[5] A. Dimofte, G. A. Bucagu, **T. M. Ingolfsson**, et al.,"CEReBrO: Compact Encoder for Representations of Brain Oscillations Using Efficient Alternating Attention", arXiv:2501.10885, Under review 2025

[6] **T. M. Ingolfsson**, V. J. K. Morinigo, A. Cossettini, X. Wang and L. Benini, "VowelNet: Enhancing Communication with Wearable EEG-Based Vowel Imagery, "IEEE Biomedical Circuits and Systems Conference (BioCAS), 2024

[7] L. Mei, C. Cioflan, **T. M. Ingolfsson**, et al., "Train-On-Request: An On-Device Continual Learning Workflow for Adaptive Real-World Brain Machine Interfaces," IEEE Biomedical Circuits and Systems Conference (BioCAS), 2024

[8] S. Frey, M. A. Lucchini, V. Kartsch, **T. M. Ingolfsson,** et al., "GAPses: Versatile Smart Glasses for Comfortable and Fully-Dry Acquisition and Parallel Ultra-Low-Power Processing of EEG and EOG," in IEEE Transactions on Biomedical Circuits and Systems, 2025

[9] D. Jonathan, U. Pale, A. Amirshahi, W. Cappelletti, **T. M. Ingolfsson,** et al., " SzCORE: Seizure Community Open-Source Research Evaluation framework for the validation of electroencephalography-based automated seizure detection algorithms," in Epilepsia, 2024

[10] L. Mei, **T.M. Ingolfsson,** et al., "An Ultra-Low Power Wearable BMI System With Continual Learning Capabilities," in IEEE Transactions on Biomedical Circuits and Systems, 2025

[11] L. Benfenati, **T. M. Ingolfsson,** et al., "BISeizuRe: BERT-Inspired Seizure Data Representation to Improve Epilepsy Monitoring," IEEE Engineering in Medicine and Biology Society (EMBC), 2024

[12] **T. M. Ingolfsson,** et al., "BrainFuseNet: Enhancing Wearable Seizure Detection Through EEG-PPG-Accelerometer Sensor Fusion and Efficient Edge Deployment," in IEEE Transactions on Biomedical Circuits and Systems, 2024,

[13] Y. E. Wibowo, C. Cioflan, **T. M. Ingolfsson** et al., "12 mJ Per Class On-Device Online Few-Shot Class-Incremental Learning," Design, Automation & Test in Europe Conference & Exhibition (DATE), 2024

[14] L. Schulthess, **T. M. Ingolfsson,** et al., "A leap into the future: Towards an augmented reality learning environment in ski-jumping," Current Issues in Sport Science (CISS), 2024

[15] **T. M. Ingolfsson**, et al. "Minimizing artifact-induced false-alarms for seizure detection in wearable EEG devices with gradient-boosted tree classifiers," Nature Scientific Reports, 2024.

[16] P. Busia, A. Cossettini, **T. M. Ingolfsson,** et al., "Reducing False Alarms in Wearable Seizure Detection With EEGformer: A Compact Transformer Model for MCUs," in IEEE Transactions on Biomedical Circuits and Systems, 2024

[17] S. Vostrikov, **T. M. Ingolfsson,** et al., "A Muscle Pennation Angle Estimation Framework From Raw Ultrasound Data for Wearable Biomedical Instrumentation," in IEEE Transactions on Instrumentation and Measurement, 2024

[18] L. Schulthess, **T. M. Ingolfsson**, M. Nölke, M. Magno, L. Benini and C. Leitner, "Skilog: A Smart Sensor System for Performance Analysis and Biofeedback in Ski Jumping," IEEE Biomedical Circuits and Systems Conference (BioCAS), 2023

[19] **T. M. Ingolfsson,** et al., "EpiDeNet: An Energy-Efficient Approach to Seizure Detection for Embedded Systems," IEEE Biomedical Circuits and Systems Conference (BioCAS), 2023

[20] P. Busia, A. Cossettini, **T. M. Ingolfsson,** et al., "EEGformer: Transformer-Based Epilepsy Detection on Raw EEG Traces for Low-Channel-Count Wearable Continuous Monitoring Devices," IEEE Biomedical Circuits and Systems Conference (BioCAS), 2022

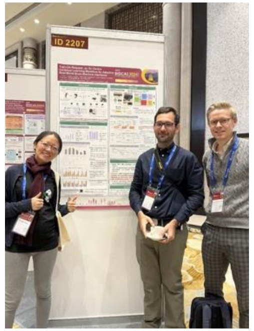
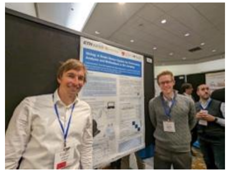
[21] **T. M. Ingolfsson**, A. Cossettini, S. Benatti and L. Benini, "Energy-Efficient Tree-Based EEG Artifact Detection", IEEE Engineering in Medicine & Biology Society (EMBC), 2022

[22] **T.M. Ingolfsson**, et al. "Reducing neural architecture search spaces with training-free statistics and computational graph clustering." Proceedings of the 19th ACM International Conference on Computing Frontiers. 2022.

[23] **T. M. Ingolfsson** et al., "Towards Long-term Non-invasive Monitoring for Epilepsy via Wearable EEG Devices," 2021 IEEE Biomedical Circuits and Systems Conference (BioCAS), 2021

[24] **T. M. Ingolfsson**, X. Wang, M. Hersche, A. Burrello, L. Cavigelli and L. Benini, "ECG-TCN: Wearable Cardiac Arrhythmia Detection with a Temporal Convolutional Network," IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), 2021






Thank you!