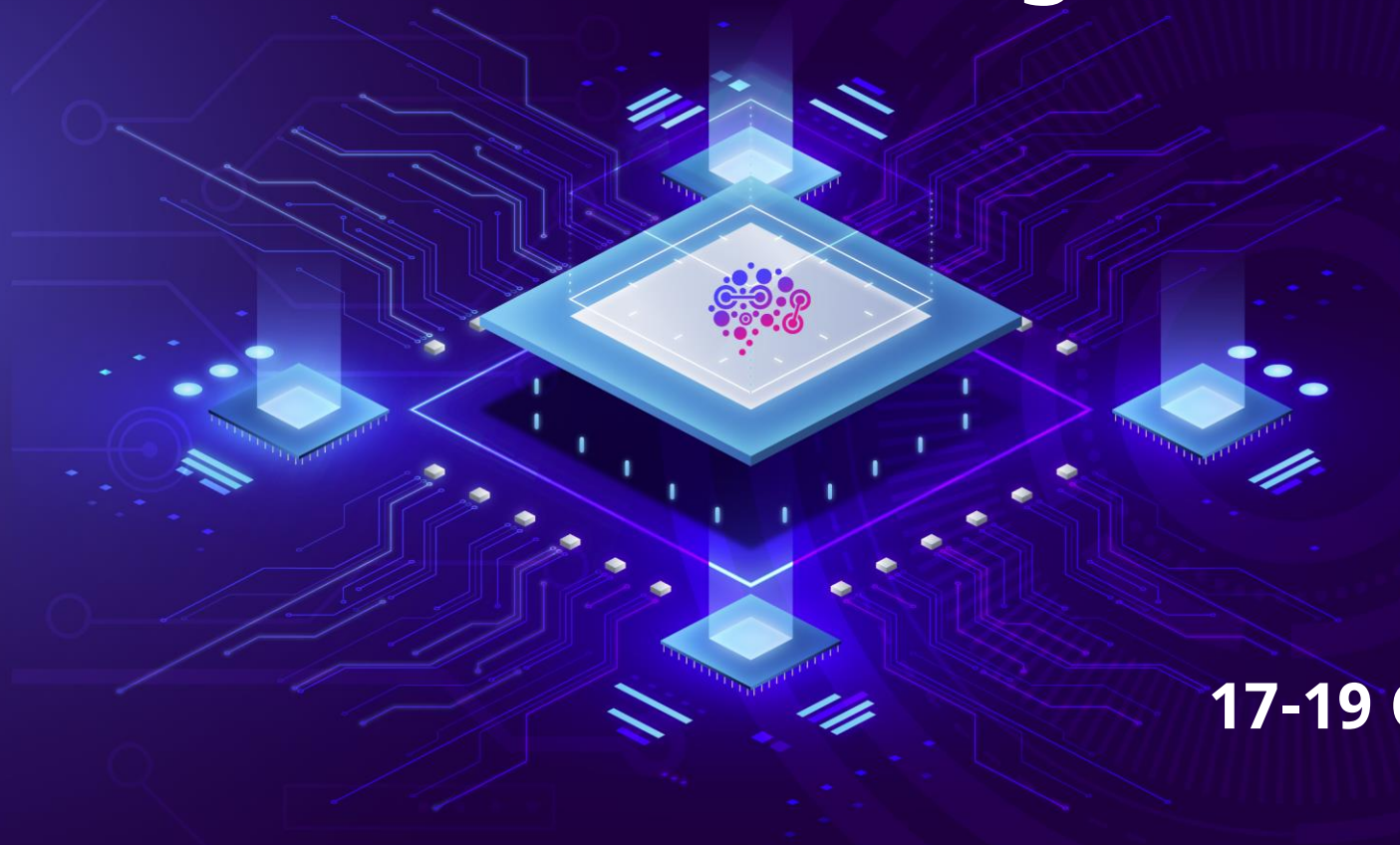
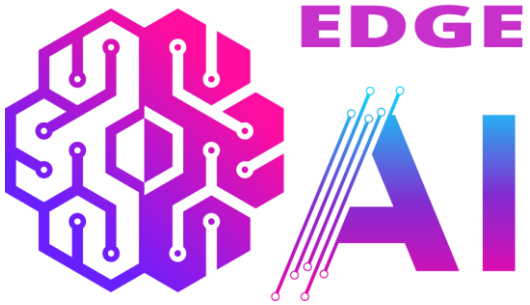
European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

European Conference on EDGE AI Technologies and Applications - EEAI



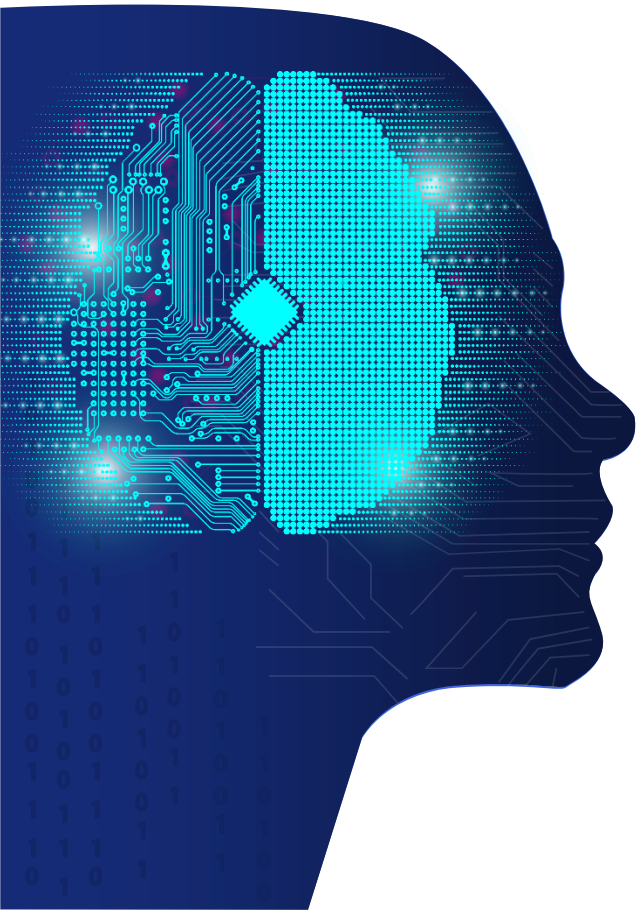
Tiny Transformers for Online Processing of Sensor Data Streams on Low Power MCUs

Paola Busia, Università degli Studi di Cagliari, Italy

17-19 October 2023 Athens, Greece

Copyright © 2023

Presentation Outline



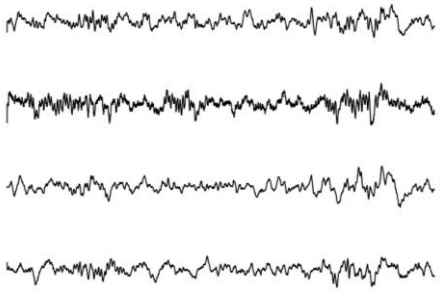
- Introduction
- State-of-the-art
- Research Goal
- Methodology
- Design and Implementation
- Experimental Results
- Future Research
- Conclusion

Copyright © 2023

Introduction
Anomaly classification on 1D-biosignals

- Strict **accuracy** standards
- **Unobtrusive** monitoring through wearable devices with minimal acquisition setups
- Robustness to **noise** degradation
- Real-time execution on **low-power** MCUs suitable for wearable monitoring
- Limited storage and computational resources

Epilepsy Monitoring

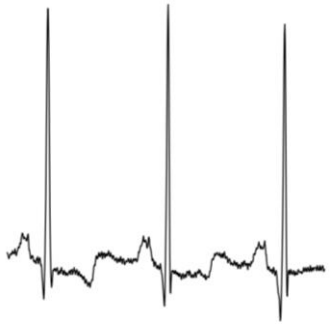


EEG signal

EEG segment CLASSIFICATION

- **N** Non-seizure segment
- **S** Seizure segment

Arrhythmia Detection



ECG signal

HEARTBEAT CLASSIFICATION

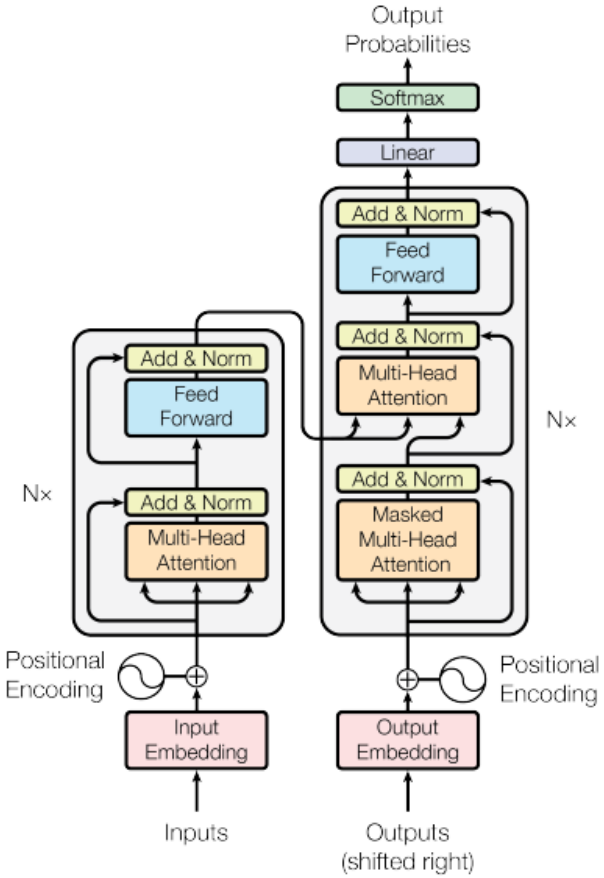
- **N** Normal
- **S** Supraventricular ectopic
- **V** Ventricular ectopic
- **F** Fusion
- **Q** Unclassified

Copyright © 2023

State-of-the-art

Transformer

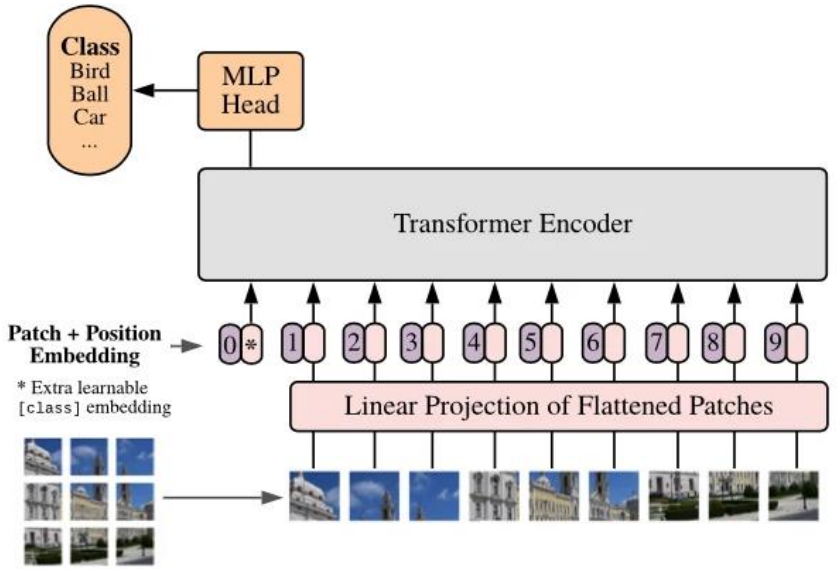
Machine Translation



Attention is all you need, A. Vaswani *et al*. NIPS 2017

Vision Transformer (ViT)

Image classification



An Image is worth 16x16 Words: Transformers for Image recognition at scale, A. Dosovitskyi *et al*. ICLR 2021

Signal Processing

Seizure detection

- Access to full acquisition setups

Seizure prediction based on transformers using scalp electroencephalogram, J. Yan *et al*. Applied Sciences 12 (2022)

Epileptic seizure prediction using deep transformer model, A. Bhattacharya *et al*. International Journal of Neural Systems 32 (2022)

Arrhythmia classification

Fusing transformer model with temporal features for ecg heartbeat classification, G. Yan *et al*. IEEE BIBM 2019

- Memory footprint and number of operations not suitable for efficient inference execution on MCUs

Copyright © 2023

Research Goal

- Design of efficient tiny transformer models for **real-time** biosignal processing on **MCUs**

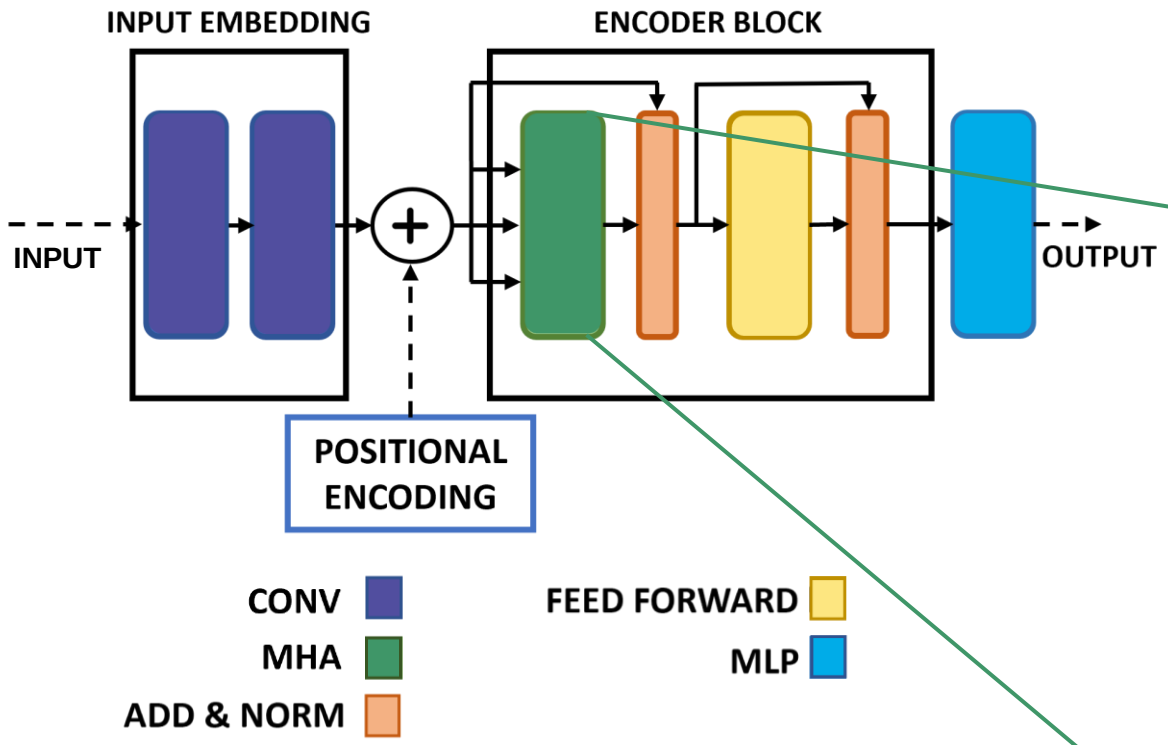
- Assessment of the classification performance on:
  - seizure detection task
  - arrhythmia detection task

- Optimization of:
  - memory **footprint** and computational complexity
  - measured **inference time**
  - measured **energy** consumption

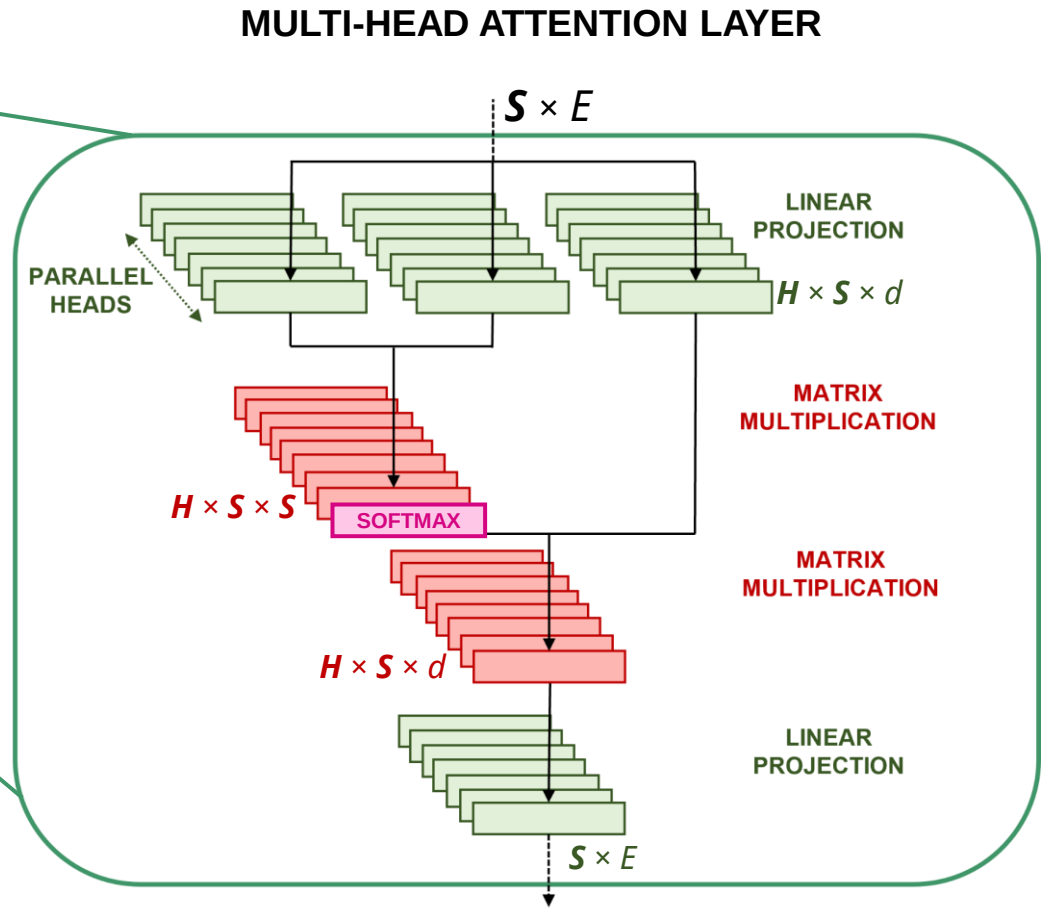
- Efficient implementation on **low-power** MCUs, for long-term monitoring

Copyright © 2023

Methodology
Transformer architecture



$$Attention(q,k,v) = Softmax\left(\frac{qk^T}{\sqrt{d}}\right) * v$$

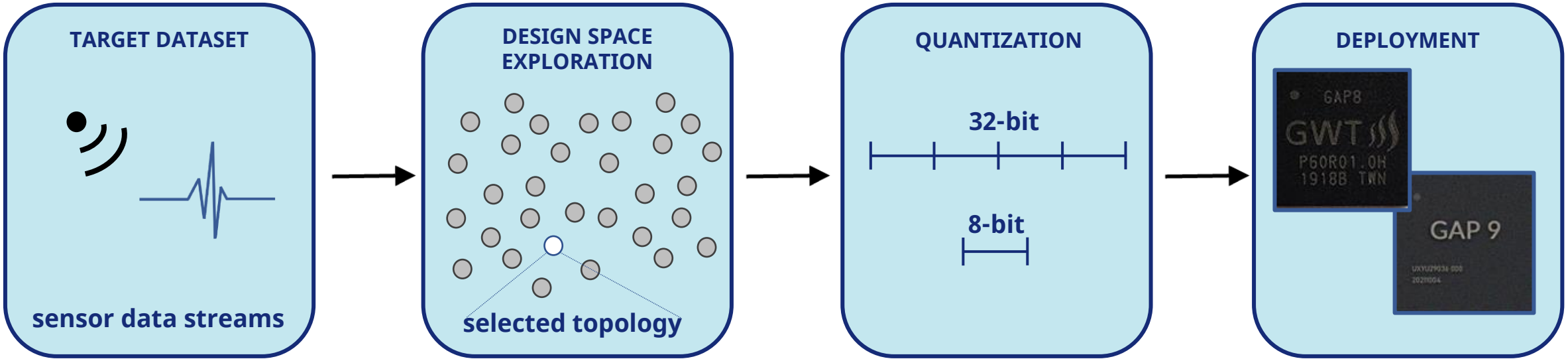


Bioformers: Embedding transformers for ultra-low power semg-based gesture recognition, A. Burrello et al., IEEE DATE 2022

EEGformer: Transformer-Based Epilepsy Detection on Raw EEG Traces for Low-Channel-Count Wearable Continuous Monitoring Devices, P. Busia et al., BioCAS 2022

Copyright © 2023

Methodology
Hardware-aware design exploration

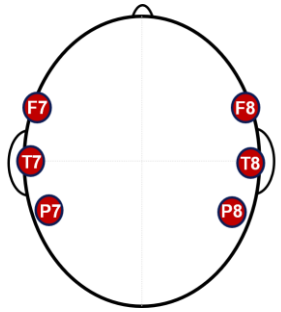


QuantLab: a Modular Framework for Training and Deploying Mixed-Precision NNs, M. Spallanzani et al.

Copyright © 2023

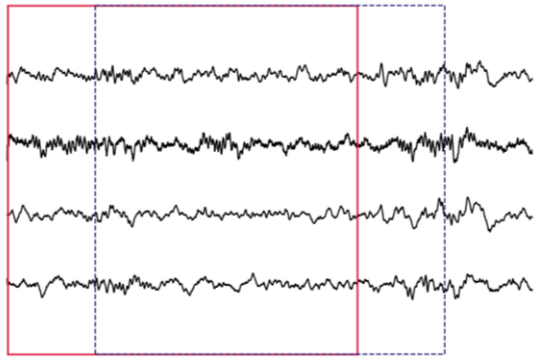
Design and Implementation
System overview

Epilepsy Monitoring

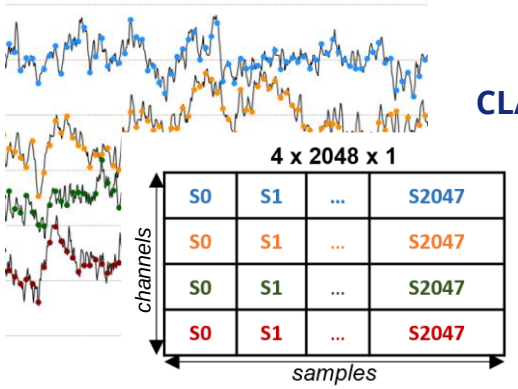


Unobtrusive acquisition

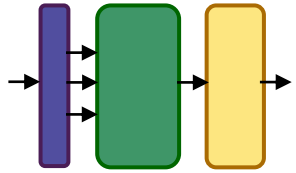
EEG TEMPORAL CHANNELS



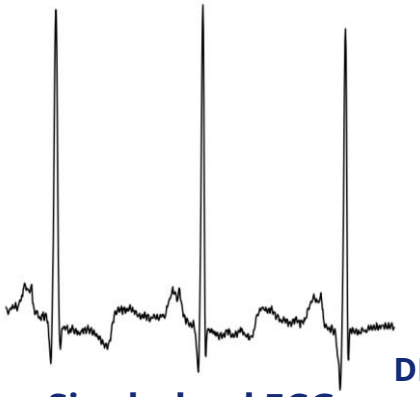
WINDOWING



CLASSIFICATION

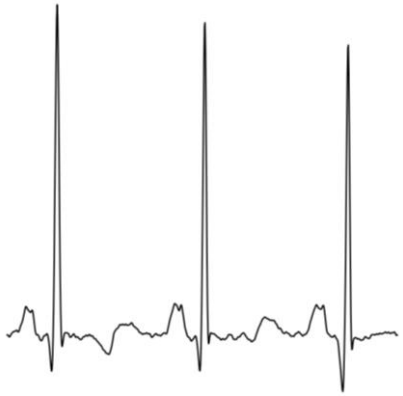


Arrhythmia Detection



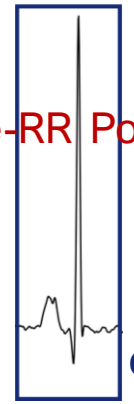
Single-lead ECG acquisition

DENOISING

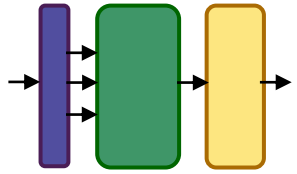


PEAK DETECTION

INPUT DATA



CLASSIFICATION



Copyright © 2023

Epilepsy Monitoring

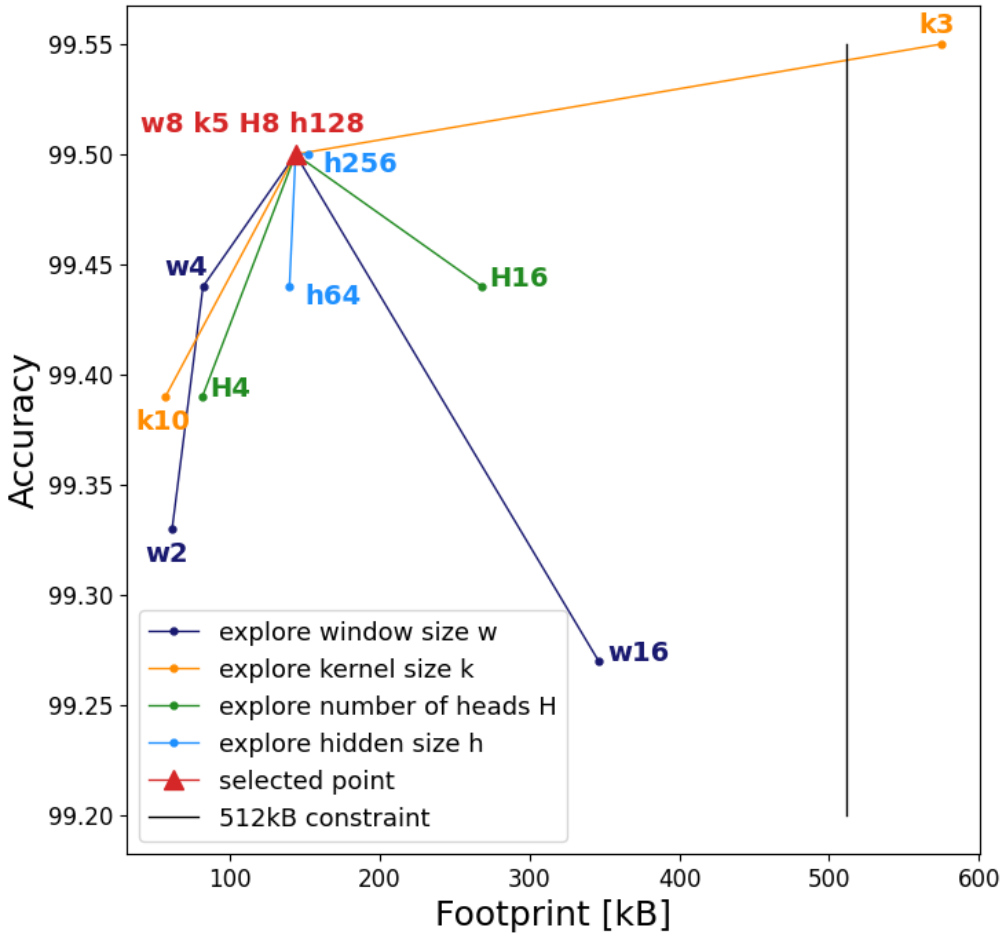
- Reference Dataset: CHB-MIT Scalp EEG dataset (256 Hz)
- **512 kB** memory constraint, based on MCUs storage resources
- Design exploration:

  - window size
  - kernel size

  ➡ Length of the sequence in MHA

  - number of heads ➡ Number of parallel executions in MHA
  - hidden layer size ➡ Workload of the FeedForward network

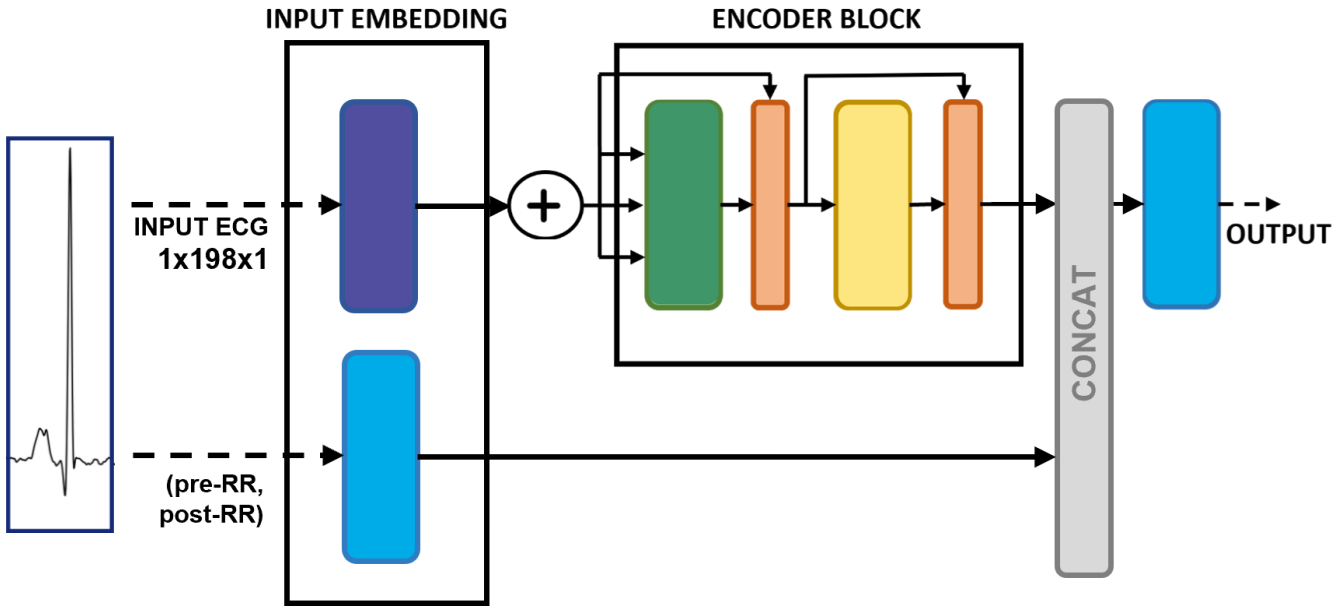
	w	K	H	h	Accuracy	Parameters	Footprint	MOPs
EEGformer	8s	5	8	128	99.5%	50.6 K	**150 kB**	14.7



Copyright © 2023

Arrhythmia detection

- Reference Dataset:
  MIT-BIH Arrhythmia dataset (360 Hz)
- **RR-intervals**, additional information for the discrimination of the arrhythmia classes
- Design exploration: **512 kB** footprint constraint



Fusing transformer model with temporal features for ecg heartbeat classification, G. Yan *et al.* IEEE BIBM 2019

	K	H	h	Accuracy	Parameters	Footprint	MOPs
ECGformer	3	8	128	99.17%	6649	**49 kB**	0.97

Copyright © 2023

Experimental Results
Classification performance assessment

Epilepsy Monitoring

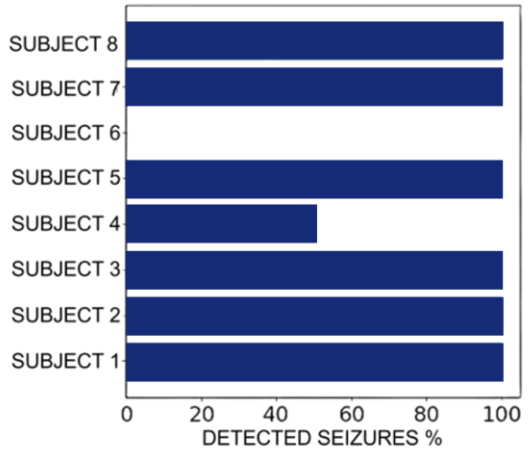
Reference Dataset: CHB-MIT Scalp EEG dataset

	Sensitivity	Specificity	FP/h	Average Latency	Detected Seizures
EEGformer	**65.5**	**99.9**	**0.15***	**15.2s**	**32/42**
AdaBoost [1] temporal ch	72	99.9	0.5	19s	38/44

* Considering consecutive FP as single events, of max duration 38s

[1] Towards Long-term Non-invasive Monitoring for Epilepsy via Wearable EEG Devices, T. M. Ingolfsson *et al*. BioCAS 2021

- Leave-one-record-out test
- **Detection of 100%** of the occurring seizure events in **6/8** patients.
- FP/h rate equal to **0 FP/h** was obtained in **87%** of tests.



EEGformer: Transformer-Based Epilepsy Detection on Raw EEG Traces for Low-Channel-Count Wearable Continuous Monitoring Devices, P. Busia et al., BioCAS 2022

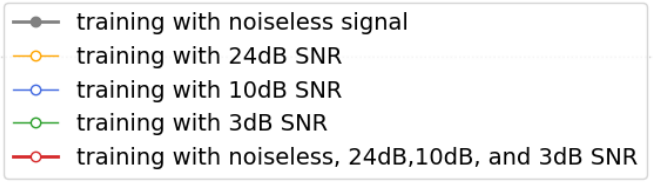
Arrhythmia Detection

Reference Dataset: MIT-BIH Arrhythmia dataset

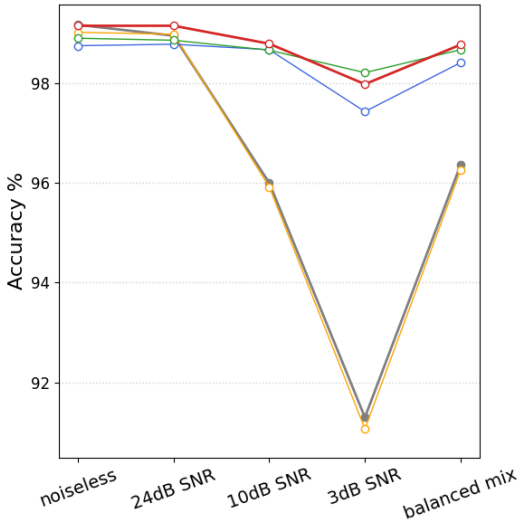
	Accuracy		Noisy Test		Number of Parameters	Memory Footprint
	32-bit	8-bit	32-bit	8-bit		
This work	99.05	**98.97**	98.65	98.36	**6649**	**49 kB**
Yan [2]	**99.62**	-	-	-	405711**	2.95 MB**

** Estimated from the paper

[2] Fusing transformer model with temporal features for ecg heartbeat classification, G. Yan *et al*. IEEE BIBM 2019



Robustness to **noise degradation** due to electrode motion artifacts, through training augmentation based on different noise levels.



Copyright © 2023

Experimental Results
On-hardware efficiency

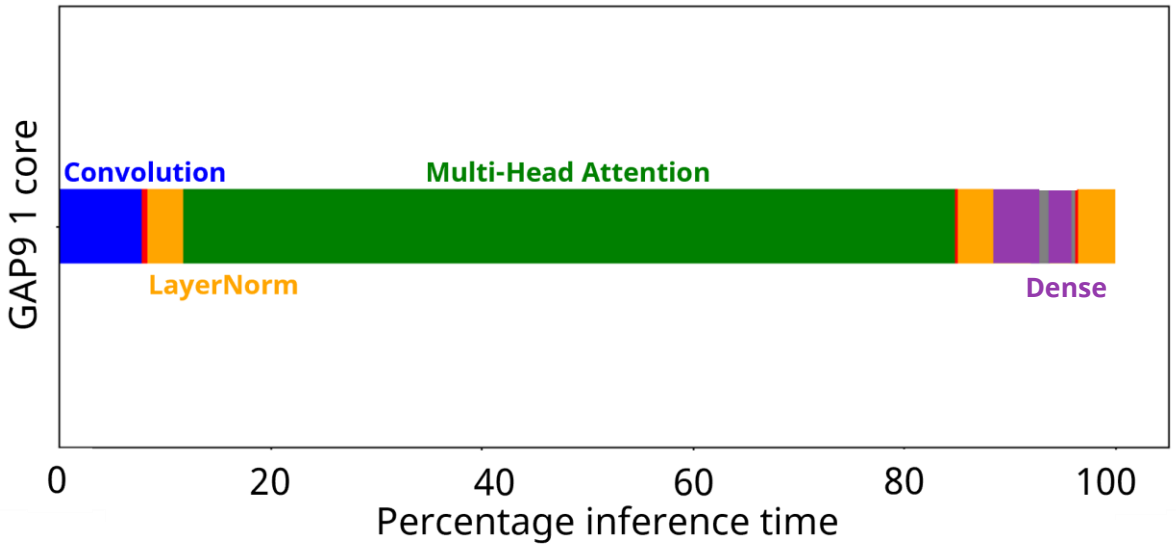
Epilepsy Monitoring

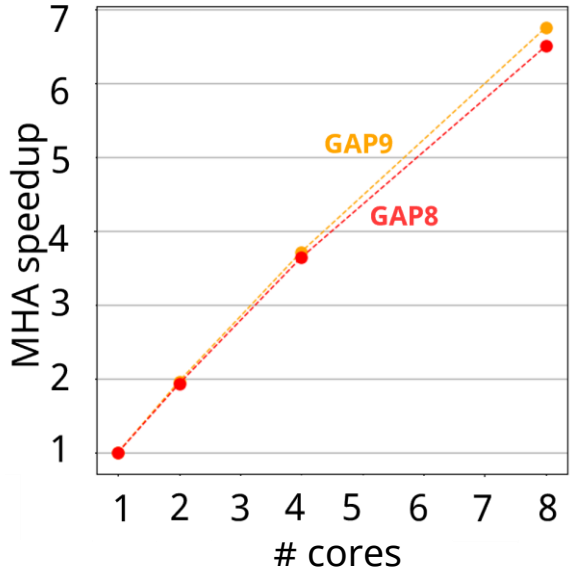
	Apollo4	GAP8		GAP9	
	1 core	1 core	8 cores	1 core	8 cores
Frequency	96 MHz	65 MHz	65 MHz	240 MHz	240 MHz
Time/ Inference	405 ms	283.9 ms	62.2 ms	72 ms	**13.7 ms**
Power	4.4 mW	10 mW	18.1 mW	12.9 mW	**22.9 mW**
Energy/ Inference	1.79 mJ	2.9 mJ	1.2 mJ	0.93 mJ	**0.31 mJ**

Arrhythmia Detection

	GAP9 8 cores
Frequency	240 MHz
Time/ Inference	**4.28 ms**
Power	**20.33 mW**
Energy/ Inference	**0.09 mJ**

- Demonstration of **real-time low-power** execution on wearable devices
- Most of the **inference time** is required for the execution of Multi-Head-Attention
- Almost **linear speedup** with the number of cores is obtained splitting MHA along the number of heads





Copyright © 2023

Future Research

- Tiny transformers for **defects detection** in manufacturing processes

  - Sensor data streams
  - Images

- Design optimization for **accuracy** and **efficiency** requirements

- Efficient implementation on MCUs

Copyright © 2023

Conclusion

- Design exploration of tiny transformer models for **real-time** biosignal processing on **MCUs**

- Classification performance comparable with the state of the art:
  - **65.5%** sensitivity with **99.9%** specificity and **0.15 FP/h** on the seizure detection task
  - **98.97%** accuracy on the arrhythmia detection task

- Quantization to 8-bit precision with **limited accuracy drop**

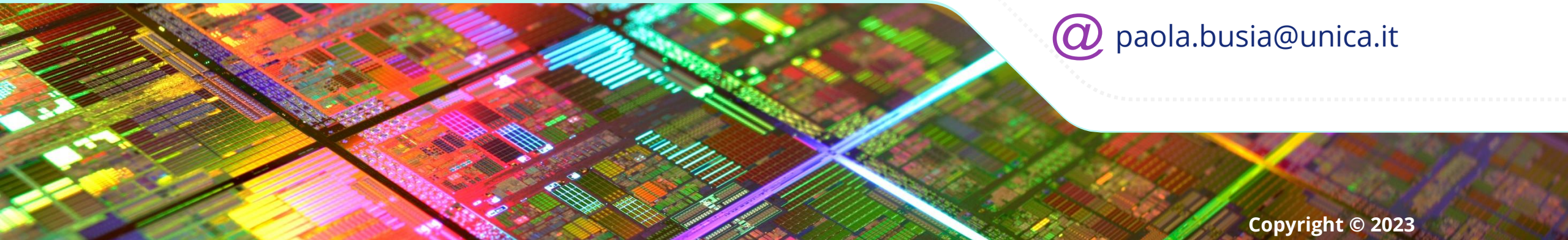
- Energy efficiency for wearable deployment evaluated on the GAP9 processor
  - **13.7 ms** inference time and **0.31 mJ/inference** on the seizure detection task
  - **4.28 ms** inference time and **0.09 mJ/inference** on the arrhythmia detection task

Copyright © 2023

EDGE AI

Thank You

For your attention



@ paola.busia@unica.it

Copyright © 2023