tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

"Advancing Medical Imaging Analysis with Multi-task and Hardware-Efficient Neural Architecture Search"

Hadjer Benmeziane – Visiting Researcher, IBM Research Europe

January 23, 2024

T I N Y ML

www.tinyML.org

TINY ML TALKS webcast

Thank you, **tinyML Strategic Partners**, for committing to take tinyML to the next Level, together

AIZIP
ARDUINO
arm
brainchip
EDGE IMPULSE

embedUR
EFFICIENT
GREENWAVES TECHNOLOGIES
Grovety Inc.
Himax

imagimob
infineon
innatera
Nota AI
NXP
P&G

PIMIC
POLYN TECHNOLOGY
Qualcomm
RENESAS
Schneider Electric
SensiML

SONY
SILICON LABS
ST life.augmented
synaptics
SYNTIANT
TDK

Executive Strategic Partners

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

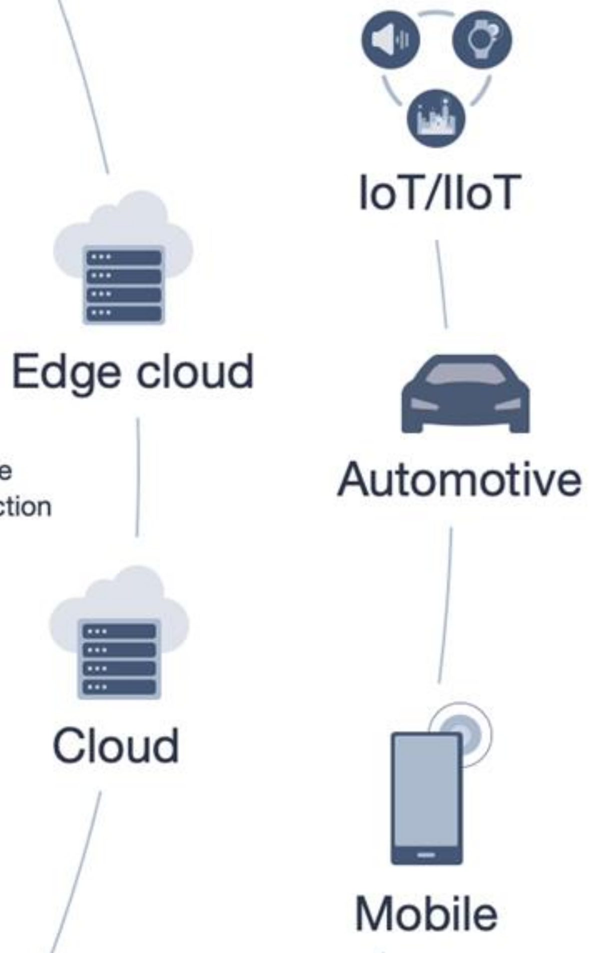
Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making



Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com

TINY ML TALKS webcast

Platinum Strategic Partners

embedUR

DEPLOY VISION AI AT THE EDGE AT SCALE

Gold Strategic Partners

Build the Future of tinyML

on arm

TINY ML TALKS webcast

EDGE IMPULSE

The Leading Development Platform for Edge ML

edgeimpulse.com

Driving decarbonization and digitalization. Together.

Infineon serving all target markets as
Leader in Power Systems and IoT

www.infineon.com

innatera

NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE

www.innatera.com

Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.

renesas.com

BIG IDEAS FOR EVERY SPACE

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy

www.st.com/ai

synaptics

ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com

Silver Strategic Partners

Join Growing tinyML Communities:

18.9k members in
49 Groups in 41 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

4.2k members
&
14.5k followers

The tinyML Community

https://www.linkedin.com/groups/13694488/

TINY ML TALKS webcast

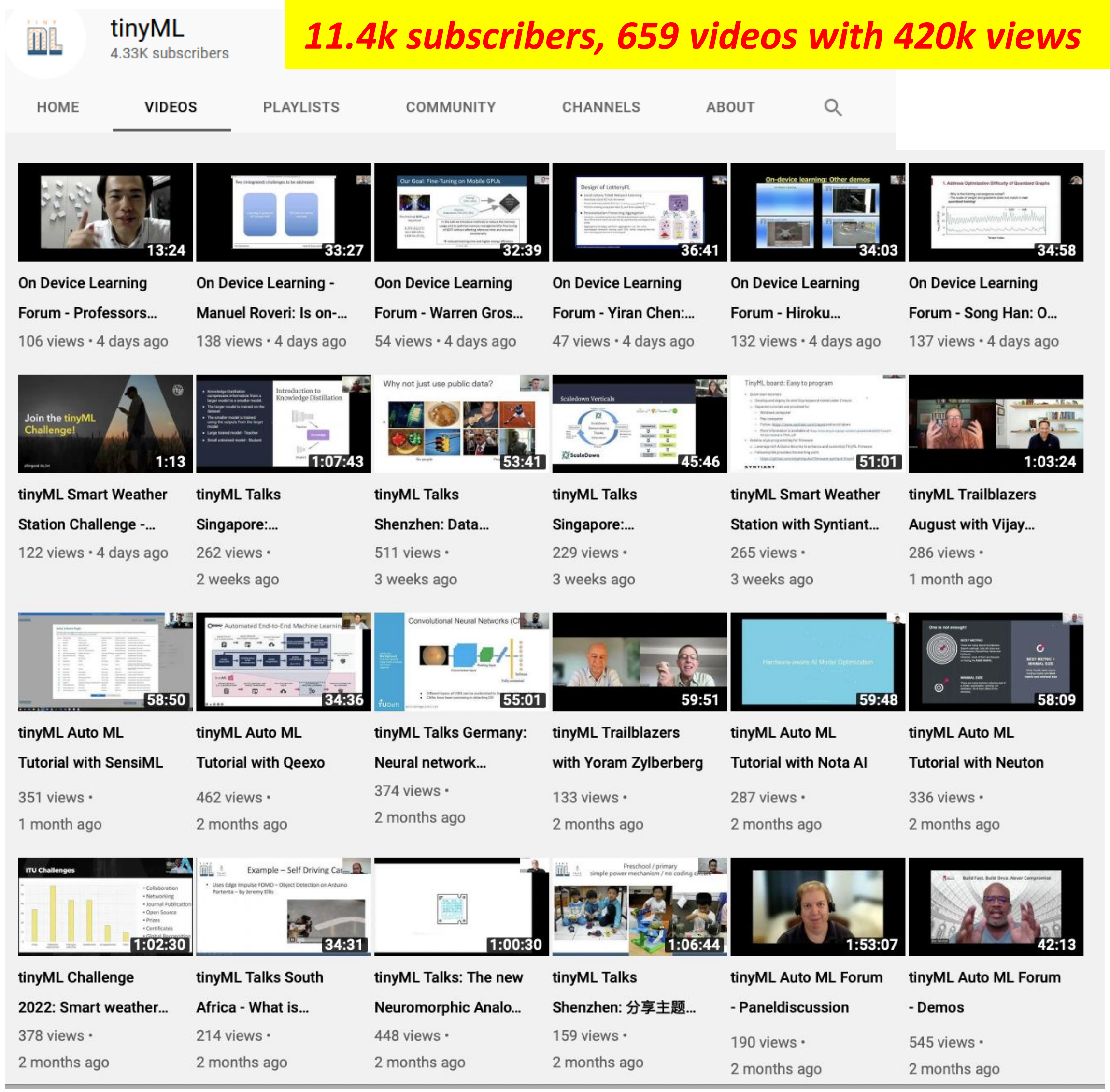
Subscribe to tinyML YouTube Channel for updates and notifications

(including this video)

<u>www.youtube.com/tinyML</u>

YouTube

11.4k subscribers, 659 videos with 420k views



tinyML Research Symposium
April 22, 2024
Call for Papers

WEVOLVER

https://www.wevolver.com/article/2023-edge-ai-technology-report

TINY ML FOUNDATION

Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums youtube.com/tinyml

Please use the Q&A window for your questions

Hadjer Benmeziane

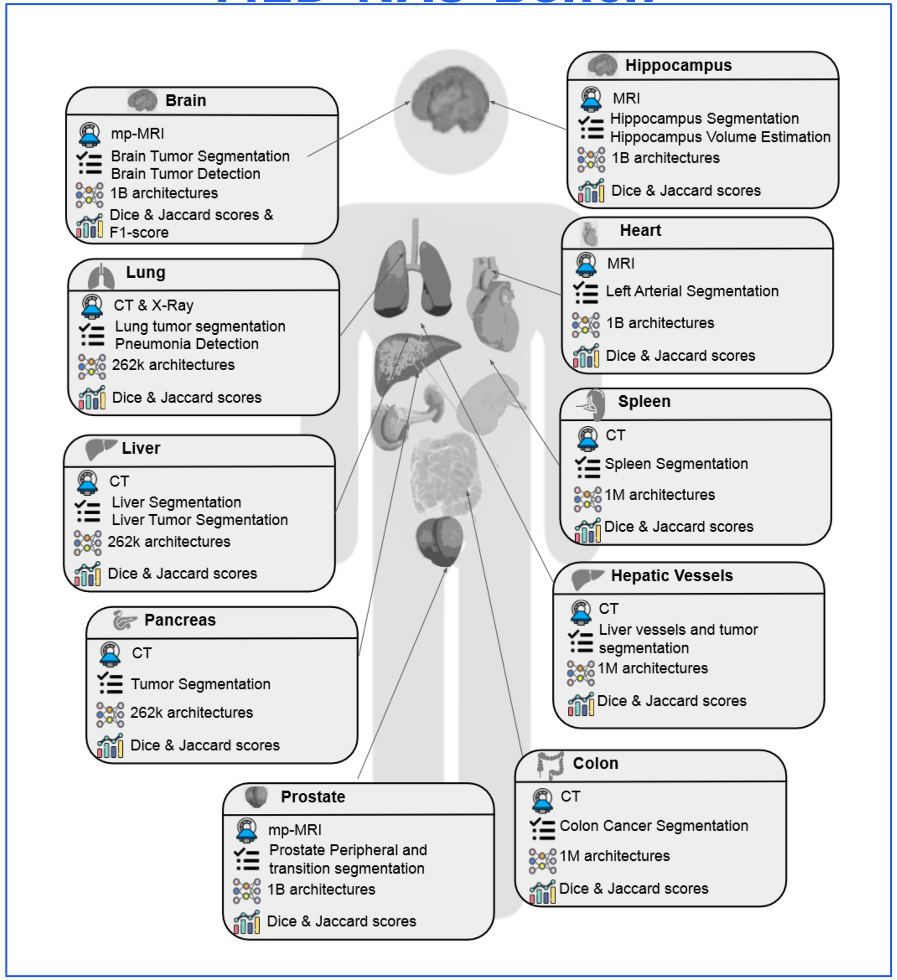
Dr. Hadjer Benmeziane is a visiting researcher at IBM Research Europe, specializing in hardware-aware neural architecture search for emerging AI accelerators such as analog in-memory computing. She received her PhD from Université Polytechnique des Hauts-de-France in August 2023, following her Master's and Engineering degree in Computer Science from Ecole Supérieure d'Informatique, Algiers, Algeria. Her work on Analog Neural Architecture Search received the prestigious IEEE open source science award and best paper award at IEEE Services Computing 2023 Symposium. Her research focuses on making hardware-aware neural architecture search more efficient, flexible and practical.

Advancing Medical Imaging Analysis with Multi-task and Hardware-Efficient Neural Architecture Search

Hadjer Benmeziane

23/01/2024

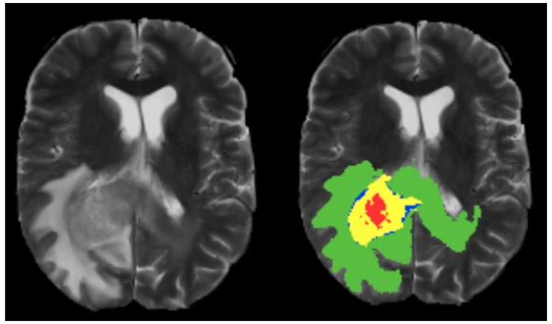
MED-NAS-Bench



Presentation Content

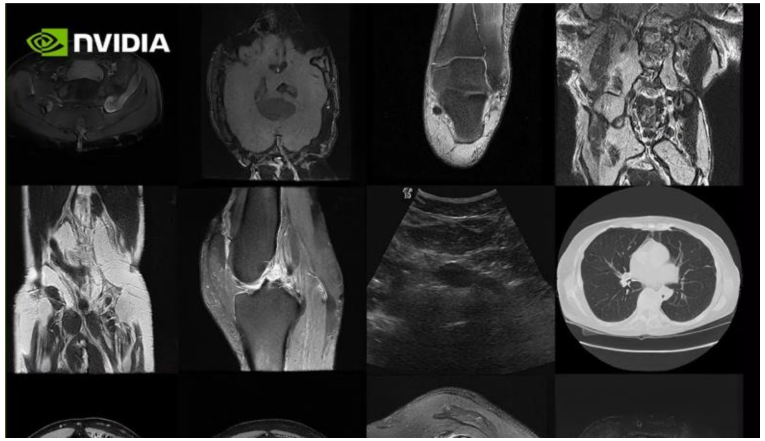
Context & Motivation

Medical Imaging Segmentation



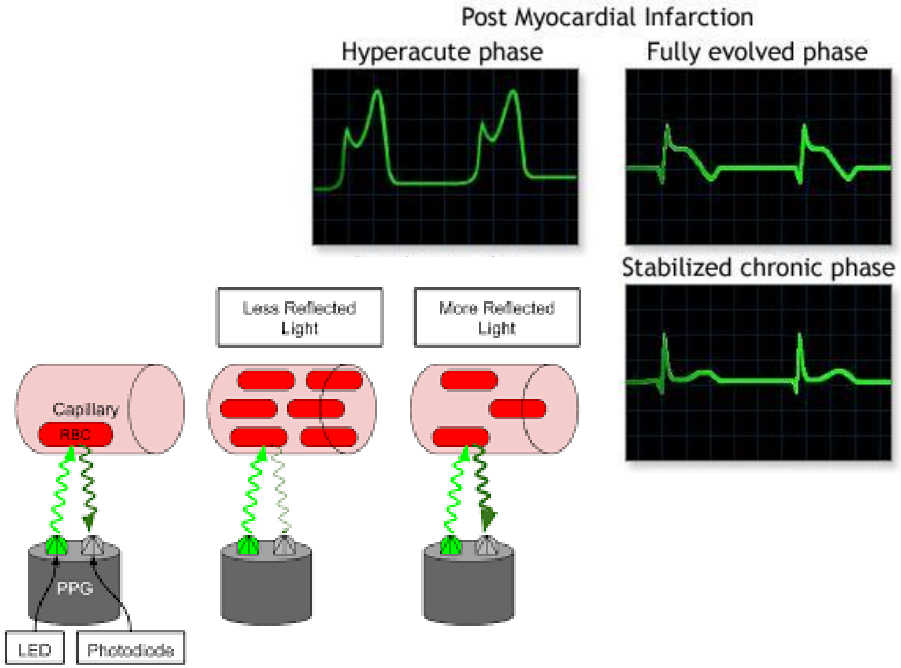
(Isensee, Fabian, et al.2021)

Generative AI for Medical Tasks



(Zhang, Peng, et al. 2023)

Medical Tasks on Wearable devices



(Banerjee, Amit, et al. 2020)

Context & Motivation

- A high number of deep learning algorithms are now FDA-approved.
- Many are targeting the fields of Radiology, Cardiology and Internal Medicine/General.
- Growing need in automating the design the these algorithms to enable multi-task benefits.

#	Name of device or algorithm	Name of parent company	Short description	FDA approval number	Type of FDA approval	Mention of algorithm in announcement	Date	Medical specialty	Secondary medical specialty
1	Arterys Cardio DL	Arterys Inc.	Software analyzing cardiovascular images from MR	K163253	510(k) premarket notification	Deep learning	2016 11	Radiology	Cardiology
2	EnsoSleep	EnsoData, Inc.	Diagnosis of sleep disorders	K162627	510(k) premarket notification	Automated algorithm	2017 03	Neurology	
3	Arterys Oncology DL	Arterys Inc.	Medical diagnostic application	K173542	510(k) premarket notification	Deep learning	2017 11	Radiology	Oncology
4	Idx	IDx LLC.	Detection of diabetic retinopathy	DEN180001	de novo pathway	AI	2018 01	Ophthalmology	
5	ContaCT	Viz.AI.	Stroke detection on CT	DEN170073	de novo pathway	AI	2018 02	Radiology	Neurology

List of FDA-approved Medical Algorithms

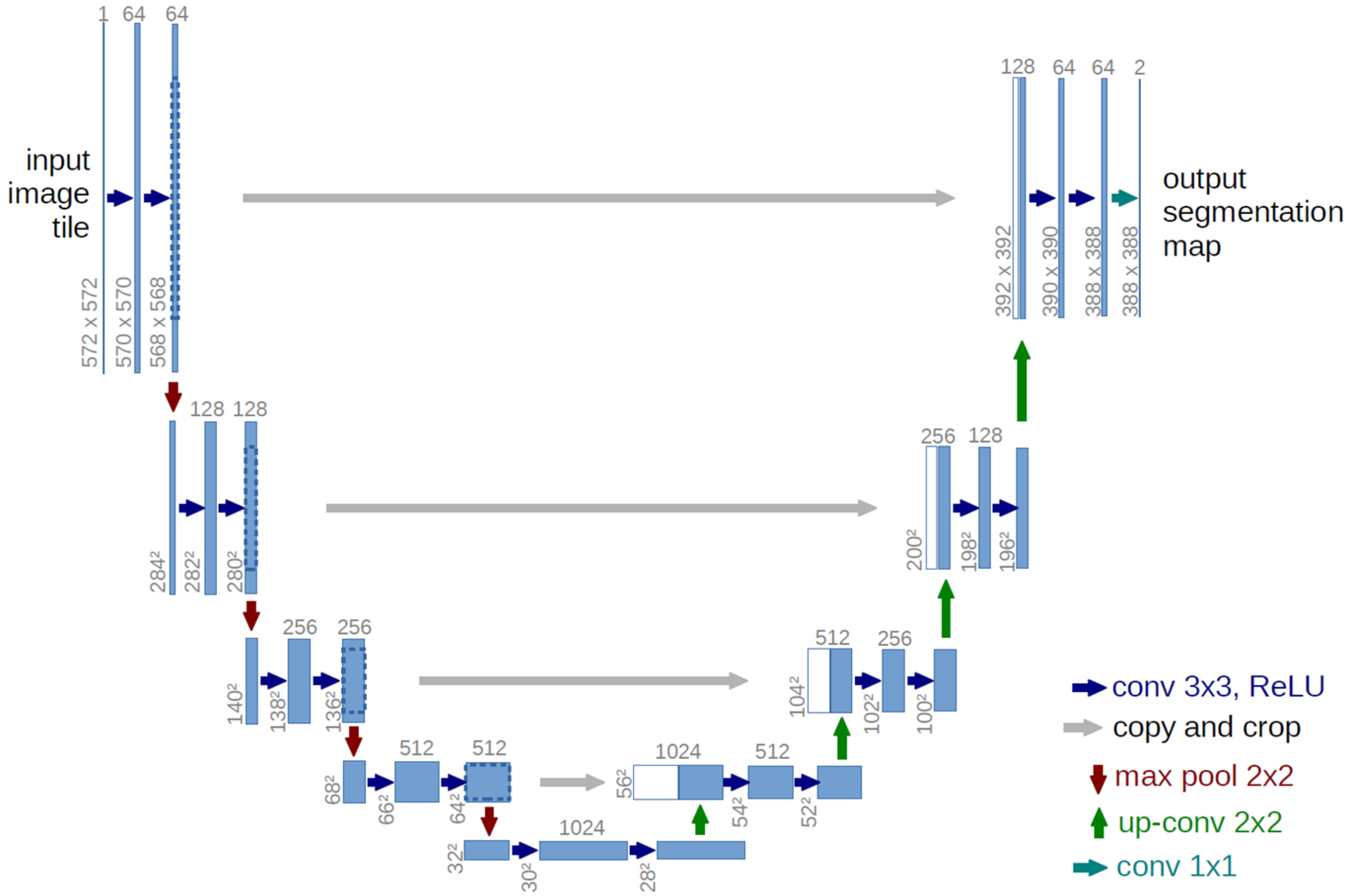
(Benjamens, Stan, et al. 2020)

Context & motivation

★ Deep Learning models are more accurate and automate many radiologist tasks.

❑ Urgent need for efficient AI solutions in fast-paced hospital environments.
❑ Current U-net like architectures are slow and large.
❑ Other constraints are more critical in medical tasks such as robustness, certainty, interpretability.

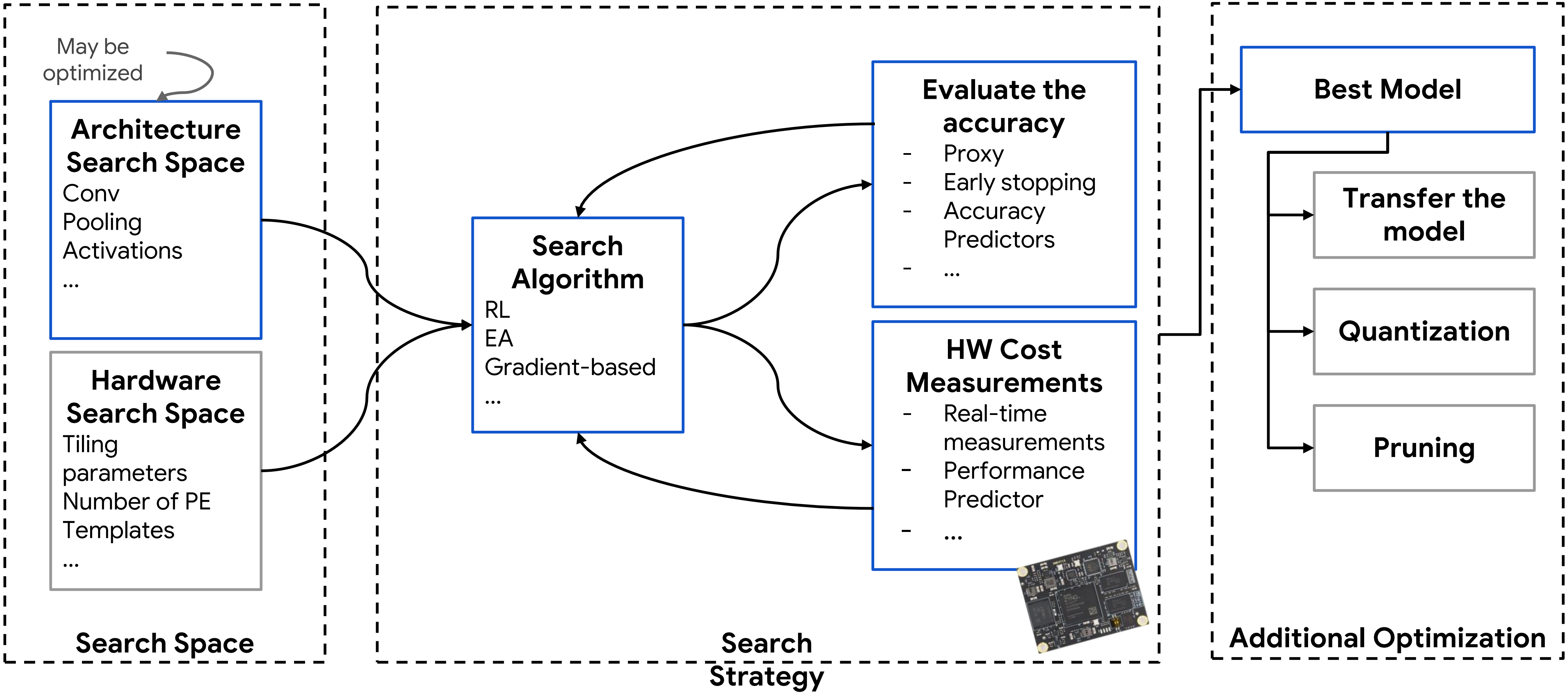
Background: Deep Learning Architectures



U-Net Architecture

(Long, Shelhamer, et al. 2014)

Background: Hardware-aware Neural Architecture Search



- Optional
- Found in all HW-NAS

Medical Imaging Analysis with HW-NAS

MED-NAS-Bench

Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

- The absence of standardized benchmarks for HW-NAS in medical imaging makes it difficult to compare and validate the effectiveness of different approaches.
- Medical imaging data often requires specialized preprocessing, such as noise reduction or contrast enhancement, which can influence the effectiveness of the derived architectures.
- Ensuring generalization across diverse patient populations and imaging conditions is challenging, given the high variability in medical images.

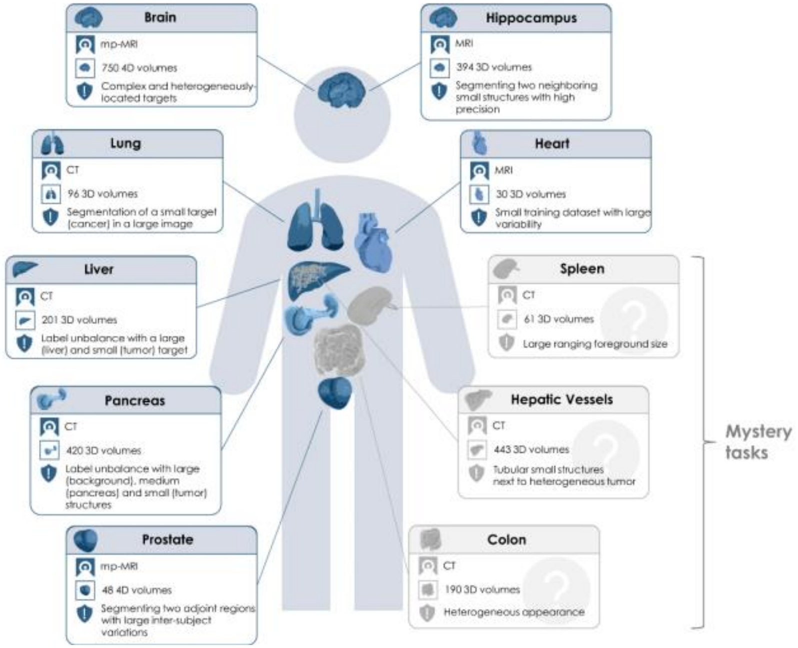
Article | Open Access | Published: 15 July 2022

The Medical Segmentation Decathlon

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, ... M. Jorge Cardoso + Show authors

Nature Communications **13**, Article number: 4128 (2022) | Cite this article

20k Accesses | **37** Citations | **51** Altmetric | Metrics

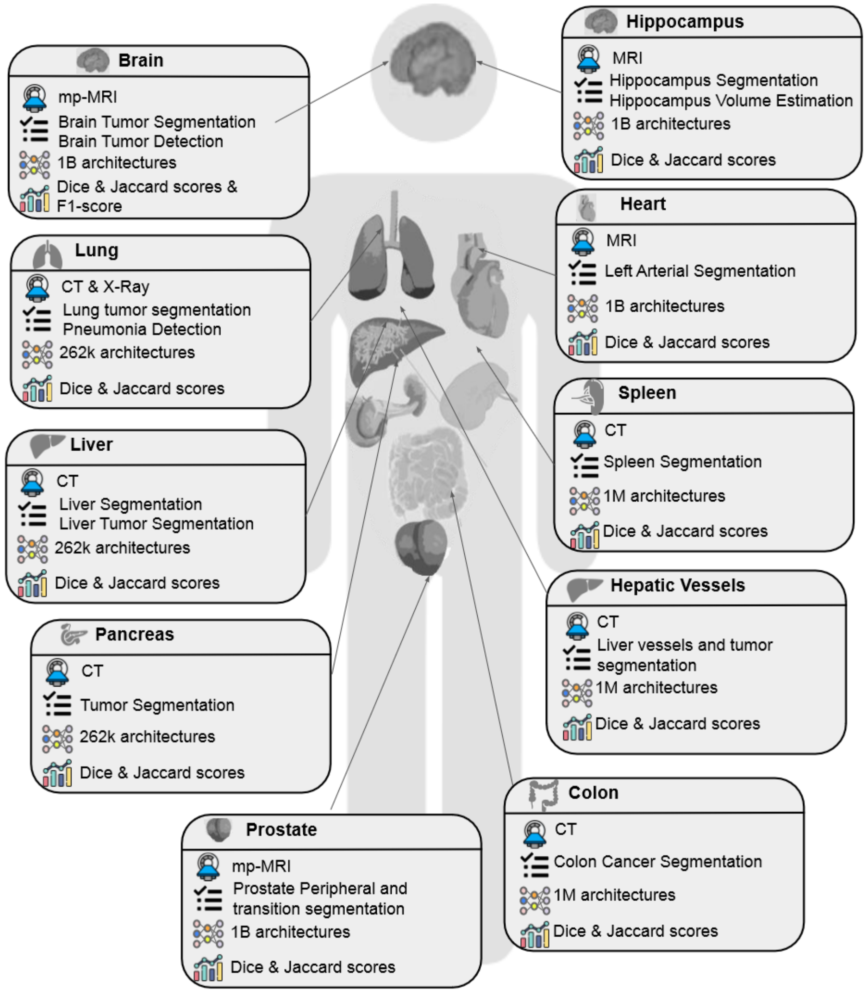


Antonelli, Michela, et al. "The medical segmentation decathlon." *Nature communications* 13.1 (2022): 4128.

Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

We propose a NAS benchmark for medical AI imaging analysis.

- The benchmark targets 11 tasks.
  - Medical Segmentation Decathlon[1]
  - NIH Chest X-Ray Dataset[2]
- The benchmark architectures are represented within a **supernetwork**.
- The supernetwork is a **U-Net like** representation.
- **Weight-sharing** methodology is used to estimate the performance scores.

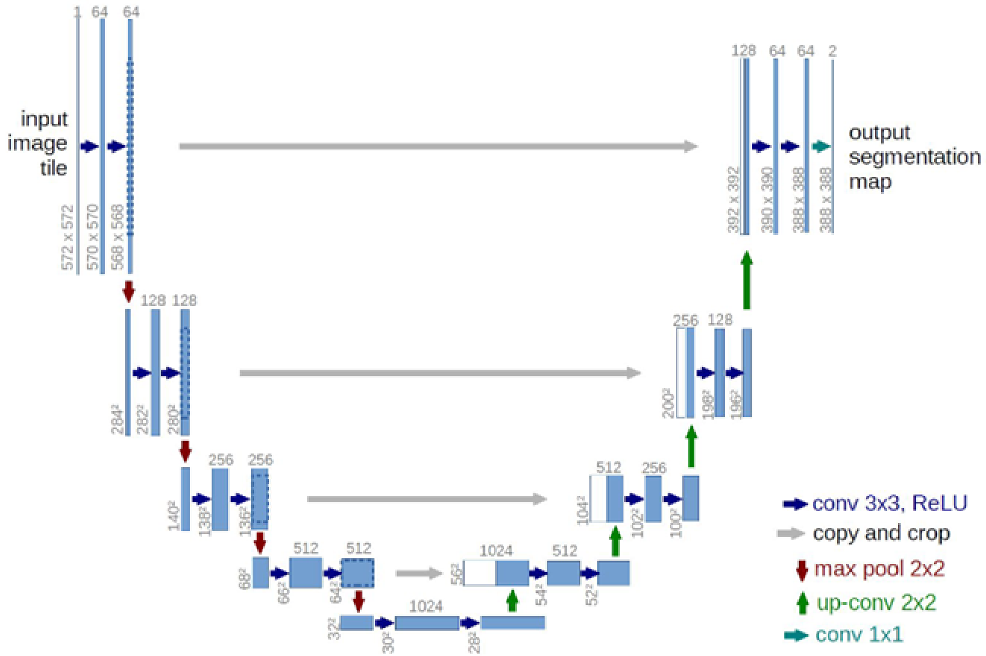


1 Antonelli, Michela, et al. "The medical segmentation decathlon." Nature communications 13.1 (2022): 4128.
2 Filice, Ross W., et al. "Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset." Journal of digital imaging 33 (2020): 490-496.

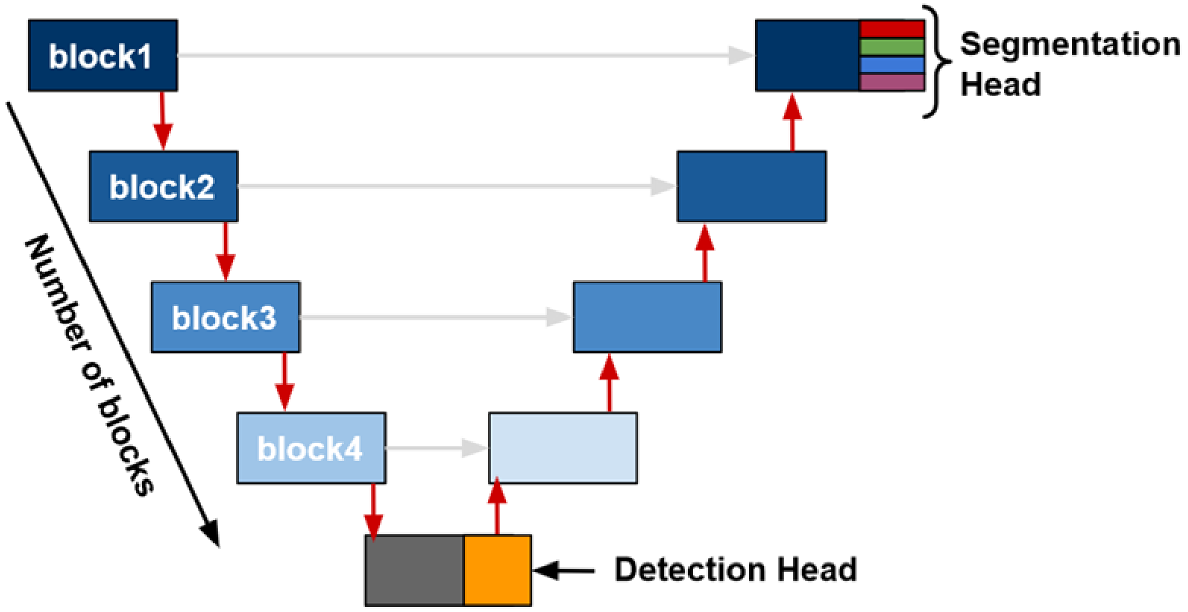
Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

- We create a supernetwork per task
- Each supernetwork is based on the U-Net architecture
- The supernetwork is created with a recursive implementation of U-Net.

(a) original U-Net Architecture



(b) Recursive U-Net Benchmark



Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

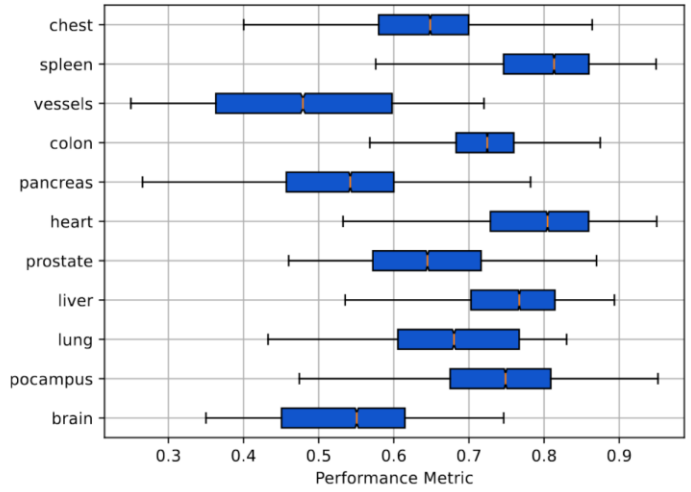
- Multiple blocks are possible per task.
- Each operation is associated with its recursive upsampling option.
- We use different 2D and 3D convolutions.

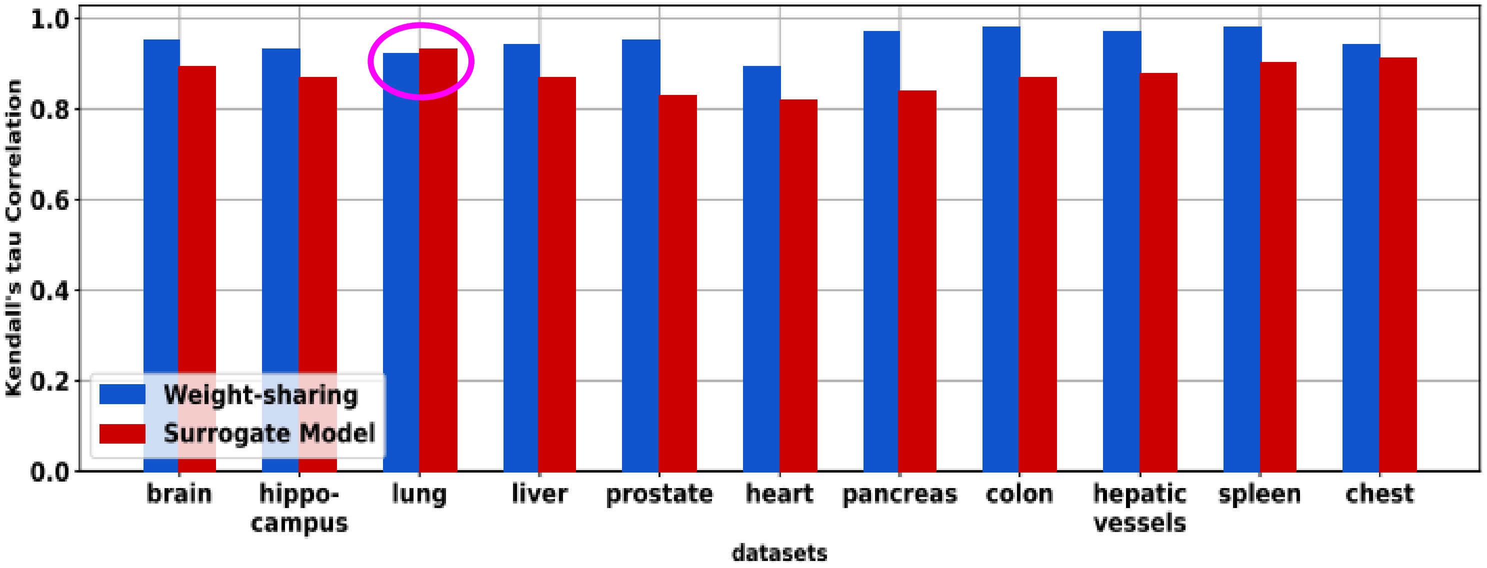
Block	Operations	Respective Upsampling block
Zero	-	-
Identity	IdentityLayer	IdentityLayer
A	LinearLayer(in, out)	LinearLayer(out, in)
B	2DConv(in, out, k=3, use_bn=false, act=relu)	T2DConv(out, in, k=2, s=2)
C	2DConv(in, out, k=3, use_bn=true, act=relu)	T2DConv(out, in, k=2, s=2)
D	2DConv(in, out, k=3, use_bn=false, act=leakyrelu)	T2DConv(out, in, k=2, s=2)
E	2DConv(in, out, k=3, use_bn=true, act=leakyrelu)	T2DConv(out, in, k=2, s=2)
F	[B, C]	T2DConv(out, in, k=2, s=2)
G	[D, E]	T2DConv(out, in, k=2, s=2)
H	3DConv(in, out, k=3, use_bn=false, act=relu)	T3DConv(out, in, k=2, s=2)
I	3DConv(in, out, k=3, use_bn=true, act=relu)	T3DConv(out, in, k=2, s=2)
J	3DConv(in, out, k=3, use_bn=false, act=leakyrelu)	T3DConv(out, in, k=2, s=2)
K	3DConv(in, out, k=3, use_bn=true, act=leakyrelu)	T3DConv(out, in, k=2, s=2)
L	[H, I]	T3DConv(out, in, k=2, s=2)
M	[J, K]	T3DConv(out, in, k=2, s=2)

Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

Ranking Evaluation

- For each task, 10k architectures were extensively trained to provide evidence that weight-sharing methodology ranking is accurate.





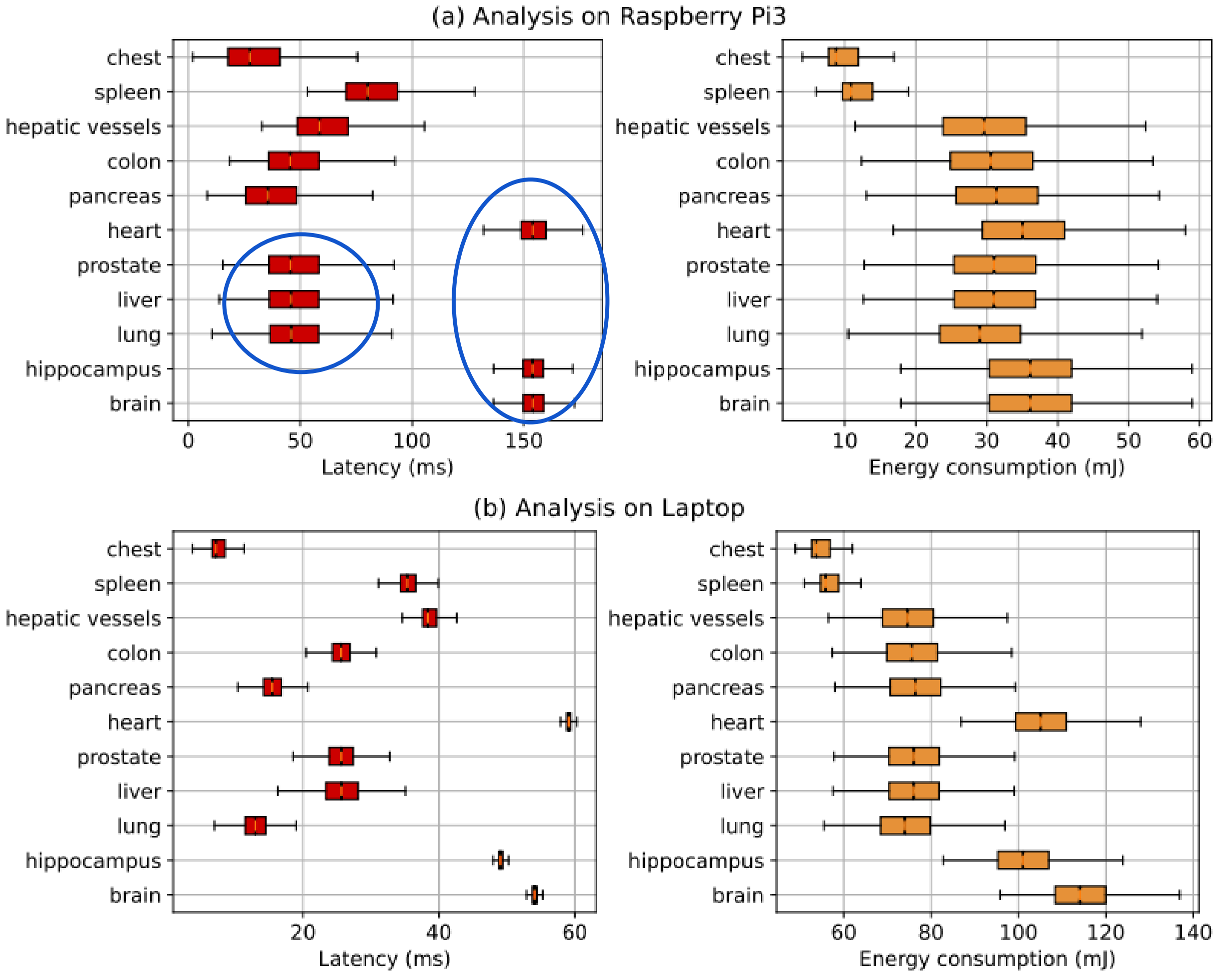
Except for lung tumor segmentation, weight-sharing method gave a better approximation of the ranking with an average of 0.85 kendall tau-b correlation.

Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

Hardware Efficiency Evaluation

- We deployed each model on two different hardware devices: Raspberry Pi3 and Laptop.
- **Except for 10K architectures in each task, the latency and energy consumption are estimated using a lookup table.**

Hardware	Processor	RAM	Storage	Operating System
Raspberry Pi 3	quad-core ARM Cortex-A53 CPU	1GB	32G	Raspbian
Laptop	AMD Ryzen 7 6800H	16GB	1T	Microsoft Windows 11

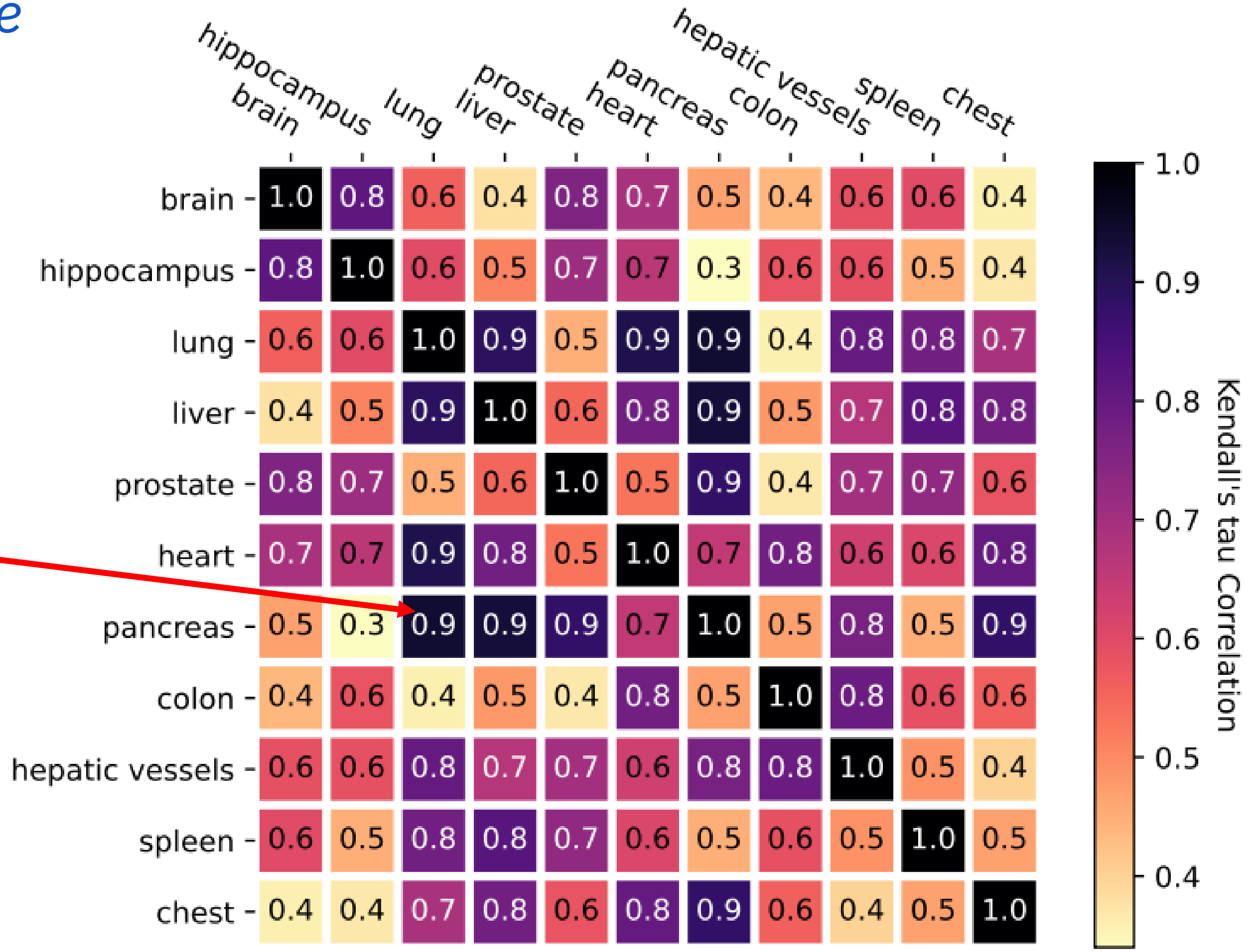


Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

Cross datasets ranking evaluation

Is there a single architecture that can be used for all medical imaging tasks ?

Similar Top Architectures



Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

two-stage differentiable neural architecture search, allowing for simultaneous cell-level and network-level optimization.

Evolutionary search on two stages: fine (operation) and coarse (connections or topology)

A two-phase search includes a differentiable NAS to narrow down the search space, followed by a novel progressive evolutionary search.

Method	Brain		Hippocampus		Lung		Liver		Prostate	
	Dice (%)	Jc (%)	Dice (%)	Jc (%)	Dice (%)	Jc (%)	Dice (%)	Jc (%)	Dice (%)	Jc (%)
U-Net [301]	57.54	51.32	80.4	74.15	54.7	47.77	75.87	69.43	77.5	72.05
U-Net++ [302]	58.98	53.67	82.63	75.65	61.3	55.4	78.38	72.4	78.97	71.17
Att. U-Net [303]	62.4	57.4	83.24	76.14	65.7	59.54	74.6	70.28	77.78	70.74
nnU-Net [308]	61.20	54.68	89.66	85.23	69.2	62.36	84.48	78.89	82.7	78.4
C2FNAS_O [305]	61.98	55.76	88.67	82.24	70.44	63.73	83.94	79.3	81.82	74.9
EA	61.56	53.54	85.6	80.21	70.8	64.38	80.9	74.53	74.56	69.87
RS	53.5	47.4	62.45	54.85	56.7	51.21	67.88	61.1	68.4	61.94
MixSearch [304]	**65.78**	**59.65**	88.67	83.31	**81.3**	**76.08**	**87.43**	**82.23**	**86.79**	**79.79**
C2FNAS [305]	64.88	54.6	**90.54**	**83.19**	79.4	73.12	86.44	80.32	83.56	79.01
BiX-NAS [306]	63.87	56.4	89.68	84.07	75.6	68.47	87.12	82.11	81.5	75.61

Method	Heart		Pancreas		Colon		Hepatic Vessels		Spleen		Chest
	Dice (%)	Jc (%)	Dice (%)	Jc (%)	Dice (%)	Jc (%)	Dice (%)	Jc (%)	Dice (%)	Jc (%)	F1-score
U-Net	85.6	79.92	64.56	59.93	54.32	49.53	38.5	30.98	89.54	84.11	95.32
U-Net++	84.32	77.89	63.87	57.03	59.82	53.64	48.93	44.49	88.95	84.6	95.38
Att. U-Net	85.78	78.95	64.76	58.41	45.7	37.91	56.73	49.69	90.56	83.66	95.78
nnU-Net	92.77	88.39	65.9	58.47	56	51.55	66.08	59.54	96	91.34	96.8
C2FNAS_O	92.49	88.1	67.59	60.68	58.9	53.7	67.65	63.54	96.28	89.49	96.34
EA	85.76	79.84	65.3	59.64	50.8	43.23	55.78	51.07	89.76	84.29	95.78
RS	75.3	69.19	54.21	46.27	46.7	41.22	39.76	35.41	80.56	73.83	89.56
MixSearch	89.53	84.73	68.43	63.2	57.8	52.1	**71.65**	**65.12**	96.75	92.27	97.54
C2FNAS	**94.56**	**89.24**	67.82	60.18	**60.67**	**53.85**	65.42	57.46	**97.34**	**93.16**	**98.68**
BiX-NAS	94.32	86.98	**69.84**	**63.66**	57.59	50.23	66.78	61.47	96.76	90.07	96.83

Medical Imaging Analysis with HW-NAS

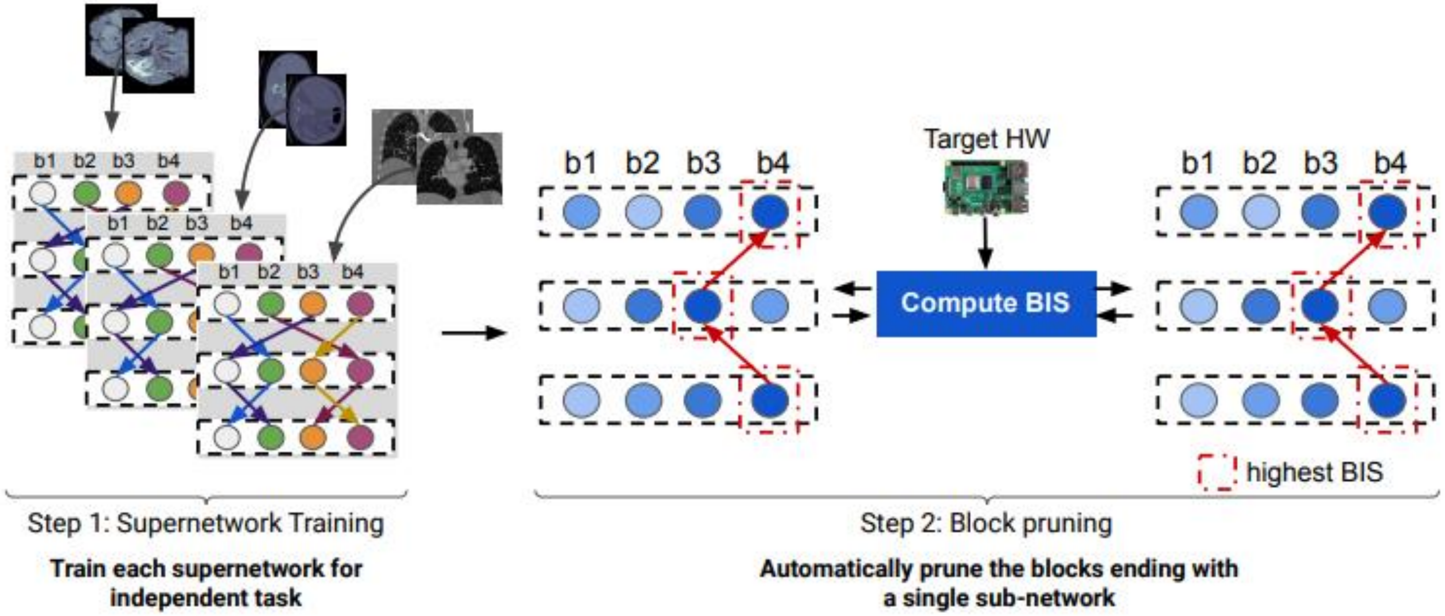
MT-MIAS

Medical Imaging Analysis with HW-NAS: MT-MIAS

Once the supernetwork is trained, we evaluate each block (layer in this case) from bottom to top.

1 Block Importance Score[4]

$$s_{l,o} = |w^{(l+1)}|^T s_{l+1}$$



★ The Pareto front is obtained by relaxing the number of selected blocks

[4]Redman, William T., et al. "AN OPERATOR THEORETIC VIEW ON PRUNING DEEP NEURAL NETWORKS." *10th International Conference on Learning Representations, ICLR 2022*. 2022.

Medical Imaging Analysis with HW-NAS: MT-MIAS

2 General Search Objective

$$\max_{o \in l} \underbrace{s_{l,o} * Avg(\alpha(A_{o \in A}))}_{\text{Validation accuracy of the operation o in layer l enhanced with OIS}} / \underbrace{Lat(o)}_{\text{HW-awareness}} + \underbrace{\sigma(s_{t,l,o} * w_o, s_{t,l,o} * w_o)}_{\text{Task-generalization ability (optional)}}$$

- alpha is the validation accuracy obtained using weight-sharing
- sigma is the cosine similarity between the weights of the same operation used in different tasks

Medical Imaging Analysis with HW-NAS: MT-MIAS

We validate MT-MIAS using three scenarios:

➔ ***MIAS Scenario:*** *Classical HW-NAS*

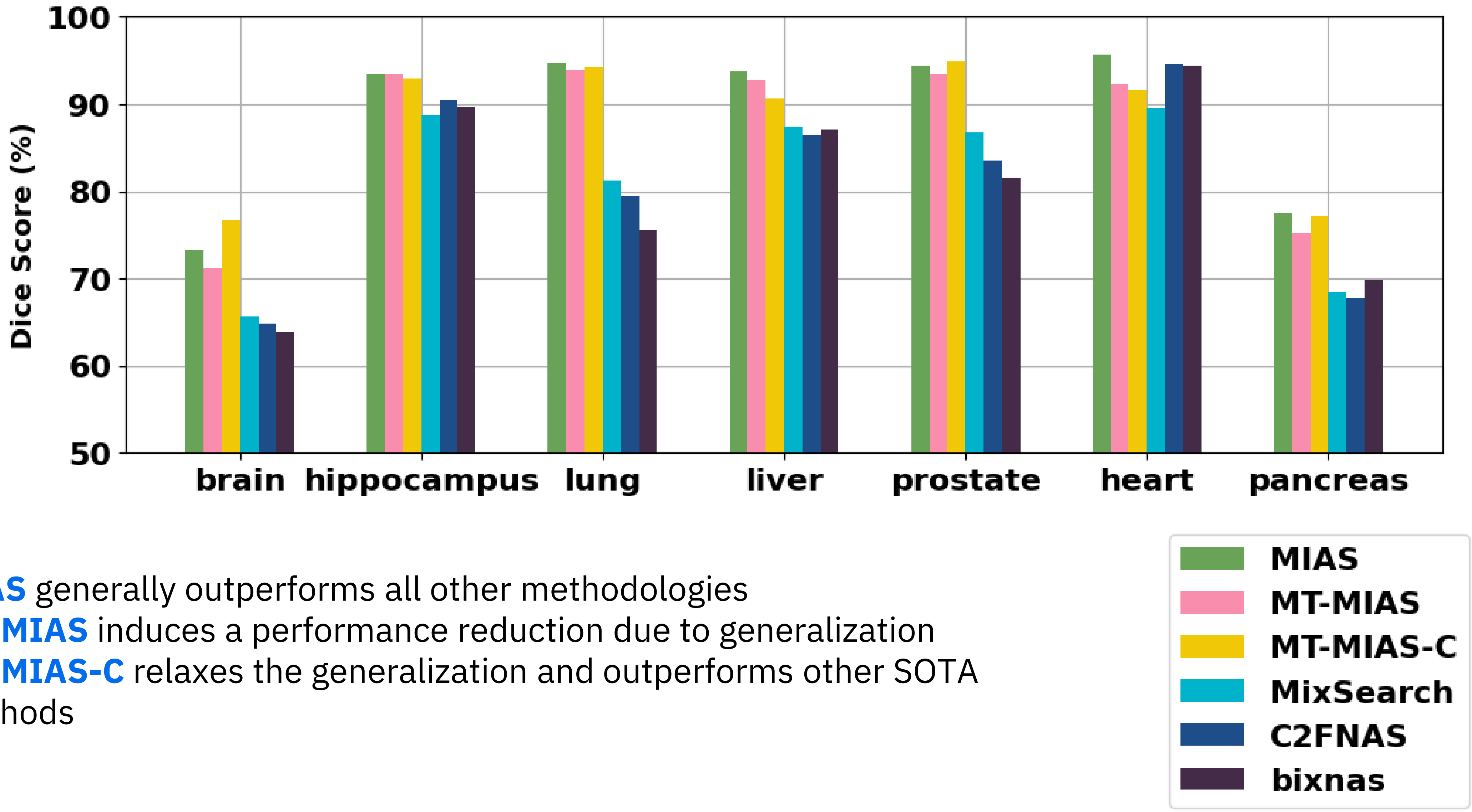
$$\max_{o \in l} s_{l,o} * Avg(\alpha(A_{o \in A}))/Lat(o)$$

➔ ***MT-MIAS Scenario:*** *Classical HW-NAS + Generalization*

$$\max_{o \in l} s_{l,o} * Avg(\alpha(A_{o \in A}))/Lat(o) + \sigma(s_{t,l,o} * w_o, s_{t,l,o} * w_o)$$

➔ ***MT-MIAS-C Scenario:*** *Relax the final number of blocks.*
The goal is to construct the smallest supernetwork that is completely deployable, but each sub-network is trained on a different task.

Medical Imaging Analysis with HW-NAS: MT-MIAS



MIAS generally outperforms all other methodologies
MT-MIAS induces a performance reduction due to generalization
MT-MIAS-C relaxes the generalization and outperforms other SOTA methods

Medical Imaging Analysis with HW-NAS: MT-MIAS

Method	Colon	Hepatic Vessels	Spleen	Chest	Hardware aware
	Dice (%)	Dice (%)	Dice (%)	F1-score	
U-Net [301]	54.32	38.5	89.54	95.32	-
U-Net++ [302]	59.82	48.93	88.95	95.38	-
Att. U-Net [303]	45.7	56.73	90.56	95.78	-
nnU-Net [308]	56	66.08	96	96.8	-
C2FNAS [305]	58.9	67.65	96.28	96.34	No
EA	50.8	55.78	89.76	95.78	No
RS	46.7	39.76	80.56	89.56	No
MixSearch [304]	57.8	71.65	96.75	97.54	No
C2FNAS [305]	60.67	65.42	97.34	98.68	No
BiX-NAS [306]	57.59	66.78	96.76	96.83	No
MT-MIAS	55.46	60.98	88.56	87.45	Yes
MT-MIAS-C	56.35	60.45	93.61	86.77	Yes
MT-MIAS (T)	63.45	70.54	95.66	93.45	Yes
MT-MIAS-C (T)	64.5	68.7	97.65	98.65	Yes

Other hardware platforms potential

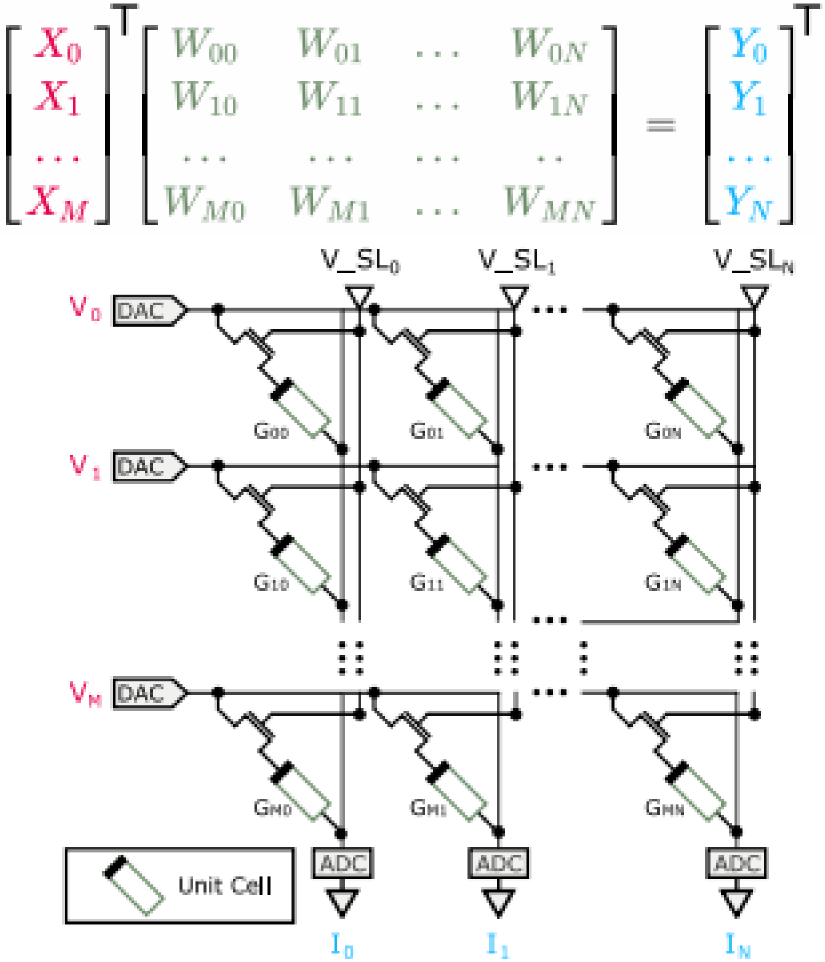
Analog in-memory computing

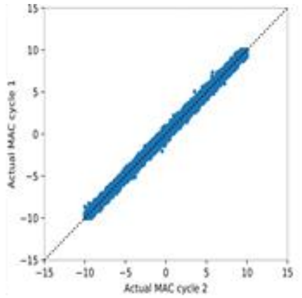
Analog In-memory Computing Potential

- Compute Vector-Matrix multiplication directly in the memory.
- **Less data movement.**
- **More energy efficient.**

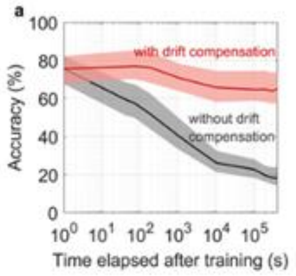
BUT!

- **Noise** and **Drift** inherent characteristics
- Accuracy drops over time
- Robust & Noise resilient architectures





Noise

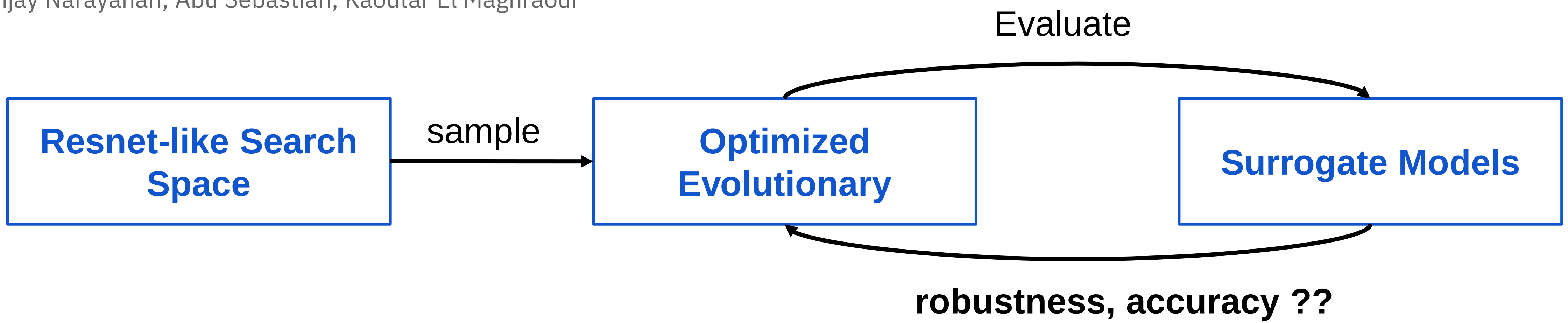


Drift

Sebastian, Abu, et al. "Memory devices and applications for in-memory computing." *Nature nanotechnology* 15.7 (2020): 529-544.

AnalogNAS: A Neural Network Design Framework for Accurate Inference with Analog In-Memory Computing *(IEEE Edge)*

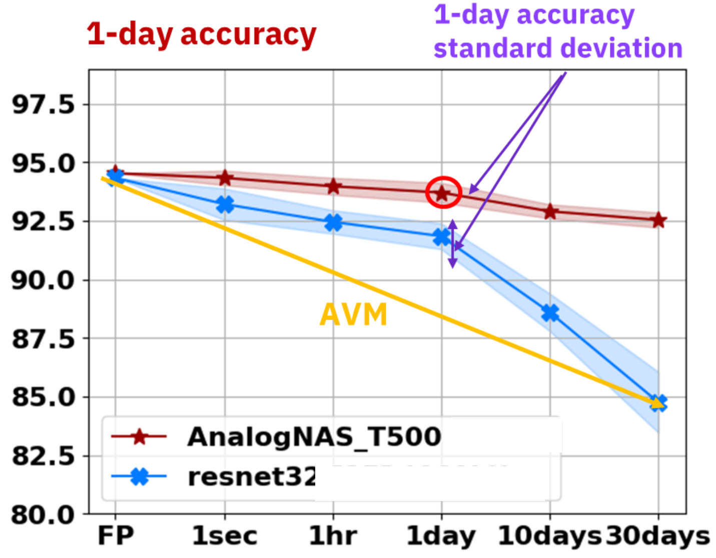
Hadjer Benmeziane, Corey Lammie, Irem Boybat, Malte Rasch, Manuel Le Gallo, Hsinyu Tsai, Ramachandran Muralidhar, Smail Niar, Ouarnoughi Hamza, Vijay Narayanan, Abu Sebastian, Kaoutar El Maghraoui



1. **The 1-day accuracy** measures the performance of an architecture on a given dataset.

1. **The Accuracy Variation over one Month (AVM)** computes the difference between the 1-month and 1-sec accuracy.

1. **The 1-day accuracy standard deviation** measures the variation of the architecture's performance across experiments.



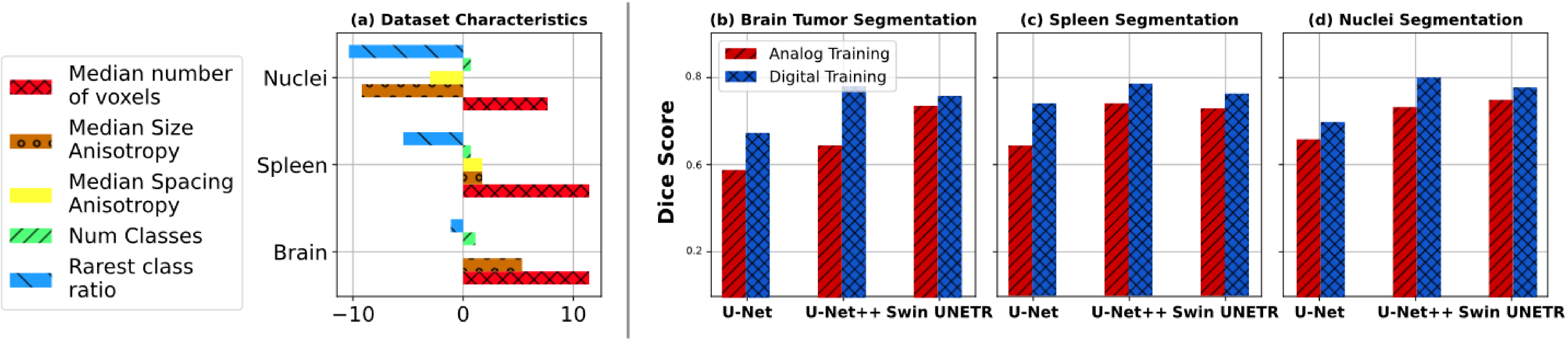
Analog In-Memory Computing with Uncertainty Quantification for Efficient Edge-based Medical Imaging Segmentation

Hadjer Benmeziane, Imane Hamzaoui, Zayneb Cherif, Kaoutar El Maghraoui

We experiments multiple medical segmentation models on different tasks, comparing analog to digital inference.

Key insights:

- ★ Transformer based models are more robust to noise injection and analog training than pyramidal architectures.
- ★ Due to additional hardware-aware training, analog inference is more reliable and certain, despite the noise.



Conclusion & Perspectives

➔ We're proud to present **MED-NAS-Bench**, a trailblazing benchmark that bridges the gap between NAS and the intricate world of medical imaging analysis.

The *MED-NAS-Bench* API represents a significant contribution to the field of medical imaging research. https://github.com/IHIaadj/med_nas_bench

➔ **MT-MIAS** encapsulates a methodology that seeks architectures optimized for holistic medical analysis, ensuring adaptability across diverse medical imaging tasks.

➔ Analog in-memory computing presents an efficient alternative to medical imaging analysis.

Perspectives:

- Increase the collected objectives and targeted hardware platforms in MED-NAS-Bench.
- Improve the API access to a sub-network in the benchmark.
- Develop a search process for automatically designing robust and noise-resilient medical architectures on analog in-memory computing.

MED-NAS-Bench

https://github.com/IHIaadj/med_nas_bench
https://theses.hal.science/tel-04224035v1/file/Benmeziane_Hadjer2.pdf
haadjer.benmeziane@gmail.com

Copyright Notice

www.tinyml.org

Copyright Notice

www.tinyml.org