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

**Imane Hamzaoui**
École nationale Supérieure d'Informatique
Algiers, Algeria
ji_hamzaoui@esi.dz

**Hadjer Benmeziane**
IBM Research Europe
8803 Rüschlikon, Switzerland
hadjer.benmeziane@ibm.com

**Zayneb Cherif**
Yorktown High school
Yorktown Heights, 10598, USA
zayneb.cherif@yorktown.org

**Kaoutar El Maghraoui**
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
kelmaghr@us.ibm.com

## Abstract

This work investigates the role of the emerging Analog In-memory computing (AIMC) paradigm in enabling Medical AI analysis and improving the certainty of these models at the edge. It contrasts AIMC's efficiency with traditional digital computing's limitations in power, speed, and scalability. Our comprehensive evaluation focuses on brain tumor analysis, spleen segmentation, and nuclei detection. The study highlights the superior robustness of isotropic architectures, which exhibit a minimal accuracy drop (0.04) in analog-aware training, compared to significant drops (up to 0.15) in pyramidal structures. Additionally, the paper emphasizes IMC's effective data pipelining, reducing latency and increasing throughput as well as the exploitation of inherent noise within AIMC, strategically harnessed to augment model certainty.

## 1 Introduction

Analog In-memory Computing (AIMC) marks a shift from traditional digital computing promising efficient and scalable processing for the rapidly growing medical data. Traditional digital systems, hindered by the Von Neumann bottleneck where data and instructions travel separately, struggle with large-scale data tasks, resulting in inherent inefficiencies. AIMC promises better efficiency and lower power use but faces challenges like susceptibility to noise, which can impact computation accuracy. In a recent study (Bonnet et al., 2023), memristor-based Bayesian neural networks (BNNs) were investigated for heartbeats classification. In contrast, our work explores AIMC's application on a wider range of medical imaging tasks, analyzing if it can effectively address healthcare needs while managing noise issues. Our primary focus lies on the algorithmic aspects, given that AIMC accelerators are still in the early stages of development (Gallo et al., 2023; Wan et al., 2022; Yin et al., 2019; Khwa et al., 2022). The code is available via this link (https://anonymous.4open.science/r/Analog_Med-B867).

## 2 Evaluating Medical Deep learning on Analog IMC

We present a comprehensive evaluation of Analog In-memory Computing (AIMC) for medical imaging, utilizing three benchmark datasets: Brain Tumor Segmentation, Spleen Segmentation, and Nuclei Detection. The study incorporates advanced architectures like U-Net (Ronneberger et al., 2015), U-Net++ (Zhou et al., 2018), and Swin Transformer (Hatamizadeh et al., 2021), trained via AIHWKIT (Rasch et al., 2021). The benchmarks and training methodologies are described in the appendix A.1 and appendixA.2. This analysis aims to assess AIMC's effectiveness in critical medical imaging tasks, highlighting its potential and capabilities in this evolving field.

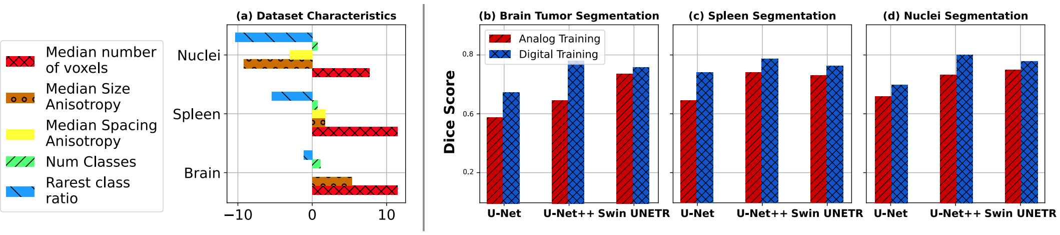


Figure 1: (a) Dissimilarity and targeted task variance. b-c-d) Noise-induced dice score drop in different medical datasets.

**Noise-resiliency in medical imaging analysis models:** We depict the diversity of the benchmarks we evaluate in Figure 1(a), highlighting various characteristics such as the size, the number of classes, etc., inspired by (Isensee et al., 2020).

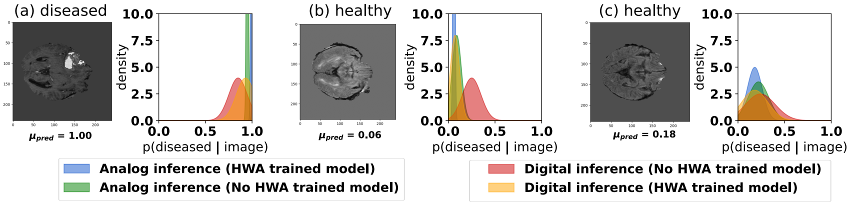
Our study on noise resilience in medical imaging models highlights two key insights. First, the pyramidal structure of U-Net models leads to increased noise vulnerability, as evidenced by dice score reductions of 0.15 and 0.22 for U-Net and U-Net++ respectively in brain segmentation tasks (see Figure 1). This vulnerability is attributed to their alternating down-sampling and up-sampling design which can amplify noise variations. In contrast, Swin-like transformer architectures exhibit remarkable noise resilience, with a negligible 0.04 performance drop as shown in Figure 1. Their isotropic design, which treats image patches consistently and lacks hierarchical convolutional operations, contributes to their enhanced stability against noise disturbances.

| Model | Avg tile utilization (%) | Avg Reuse factor | # Parameters (M) |
|---|---|---|---|
| U-Net (Ronneberger et al., 2015) | 7.42 | 6574.7 | 7.76 |
| UNet++ (Zhou et al., 2018) | 12.53 | 8721.0 | 19.6 |
| Swin UNET (Hatamizadeh et al., 2021) | 43.2 | 11540.6 | 62.19 |

Table 1: Performance metrics of state-of-the-art medical imaging models.

**Model Inference on MRIs & CT images:** In medical imaging, such as MRI and CT scans, Analog In-memory Computing (AIMC) significantly enhances data processing efficiency. Unlike traditional methods, AIMC's pipelining ability allows for rapid, parallel processing of sequential image slices, crucial for three-dimensional anatomical analysis. This approach not only improves throughput in urgent medical scenarios but also optimizes energy usage and minimizes latency between slices, making it particularly effective for volumetric data in tumor segmentation.

Figure 2: Uncertainty analysis on sample brain tumor segmentation images for UNet++.



**Certainty enhanced through noise:** While noise in computational models is often seen as detrimental, in hardware-aware training (HWA training), it inadvertently leads to more resilient models. These models, trained under controlled noise conditions, show improved noise tolerance and prediction certainty, as visualized in Figure 2. In critical healthcare applications, the certainty of model predictions is vital, as it minimizes the risk of misdiagnosis and enhances decision-making in treatment planning. This is especially evident when comparing digital and analog-trained models, such as the U-Net++ architectures.

## 3 Conclusion

In-memory computing (IMC) holds promise in refining medical imaging, with transformer structures surpassing their pyramidal counterparts in resilience to noise. Rather than being detrimental, strategic noise injection fortifies model precision, an essential aspect in healthcare. This approach mitigates overfitting and boosts confidence in diagnostics. Future efforts will focus on advancing transformer-based analog-aware architectures through the application of neural architecture search, catering to a wide array of medical imaging tasks.

## Acknowledgment

This research work was supported by Abu Dhabi National Oil Company (ADNOC), Emirates NBD, Sharjah Electricity Water & Gas Authority (SEWA), Technology Innovation Institute (TII) and GSK as the sponsors of the 4th Forum for Women in Research (QUWA): Sustaining Women's Empowerment in Research & Innovation at University of Sharjah.

We acknowledge support from the IBM Research AI Hardware Center and the Center for Computational Innovation (CCI) at Rensselaer Polytechnic Institute for computational resources on the AiMOS Supercomputer.

## References

Djohan Bonnet, Tifenn Hirtzlin, Atreya Majumdar, Thomas Dalgaty, Eduardo Esmanhotto, V. Meli, N. Castellani, Simon J. Martin, J. F. Nodin, G. Bourgeois, Jean-Michel Portal, Damien Querlioz, and E. Vianello. Bringing uncertainty quantification to the extreme-edge with memristor-based Bayesian neural networks. *Nature Communications*, 14(1), 11 2023. doi: 10.1038/s41467-023-43317-9. URL https://doi.org/10.1038/s41467-023-43317-9.

Juan C Caicedo, Allen Goodman, Kyle W. Karhohs, Beth A. Cimini, Jeanelle Ackerman, Marzieh Haghighi, Cherkeng Heng, Tim Becker, Minh Doan, Claire McQuin, Mohammad Hossein Rohban, Shantanu Singh, and Anne E. Carpenter. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature Methods*, 16(12):1247–1253, 10 2019. doi: 10.1038/s41592-019-0612-7. URL https://doi.org/10.1038/s41592-019-0612-7.

M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, Andriy Myronenko, Wentao Zhu, Yun Liu, Mingxin Zheng, Yucheng Tang, Isaac Yang, Michael Zephyr, Behrooz Hashemian, Sachidanand Alle, Mohammad Zalbagi Darestani, Charlie Budd, Marc Modat, Tom Vercauteren, Guotai Wang, Yiwen Li, Yipeng Hu, Yunguan Fu, Benjamin Gorman, Hans Johnson, Brad Genereaux, Barbaros S. Erdal, Vikash Gupta, Andres Diaz-Pinto, Andre Dourson, Lena Maier-Hein, Paul F. Jaeger, Michael Baumgartner, Jayashree Kalpathy-Cramer, Mona Flores, Justin Kirby, Lee A. D. Cooper, Holger R. Roth, Daguang Xu, David Bericat, Ralf Floca, S. Kevin Zhou, Haris Shuaib, Keyvan Farahani, Klaus H. Maier-Hein, Stephen Aylward, Prerna Dogra, Sebastien Ourselin, and Andrew Feng. Monai: An open-source framework for deep learning in healthcare, 2022.

Manuel Le Gallo, Riduan Khaddam-Aljameh, Miloš Stanisavljević, Athanasios Vasilopoulos, Benedikt Kersting, Martino Dazzi, Geethan Karunaratne, Matthias Brändli, Abhairaj Singh, Silvia M. Müller, Julian Büchel, Xavier Timoneda, Vinay Joshi, Malte J. Rasch, Urs Egger, Angelo Garofalo, Anastasios Petropoulos, Theodore Antonakopoulos, Kevin Brew, S. Choi, I. Ok, Timothy M. Philip, V. Chan, Claire Silvestre, Ishtiaq Ahsan, Nicole Saulnier, Vijaykrishnan Narayanan, Pier Andrea Francese, Evangelos Eleftheriou, and Abu Sebastian. A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference. *Nature Electronics*, 6(9):680–693, 8 2023. doi: 10.1038/s41928-023-01010-1. URL https://doi.org/10.1038/s41928-023-01010-1.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In Alessandro Crimi and Spyridon Bakas (eds.), *MICCAI*, volume 12962, pp. 272–284, 2021.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 12 2020. doi: 10.1038/s41592-020-01008-z. URL `https://doi.org/10.1038/s41592-020-01008-z`.

Win-San Khwa, Yen-Cheng Chiu, Chuan-Jia Jhang, Sheng-Po Huang, Chun-Ying Lee, Tai-Hao Wen, Fu-Chun Chang, Shao-Ming Yu, Tung-Yin Lee, and Meng-Fan Chang. A 40-nm, 2m-cell, 8b-precision, hybrid slc-mlc pcm computing-in-memory macro with 20.5 - 65.0tops/w for tiny-al edge devices. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 65, pp. 1–3, 2022. doi: 10.1109/ISSCC42614.2022.9731670.

Malte J. Rasch, Diego Moreda, Tayfun Gokmen, Manuel Le Gallo, Fabio Carta, Cindy Goldberg, Kaoutar El Maghraoui, Abu Sebastian, and Vijay Narayanan. A flexible and fast pytorch toolkit for simulating training and inference on analog crossbar arrays. *CoRR*, abs/2104.02184, 2021. URL `https://arxiv.org/abs/2104.02184`.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi (eds.), *MICCAI*, volume 9351, pp. 234–241, 2015.

Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan H. Heckers, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms, 2019.

Weier Wan, Rajkumar Kubendran, Clemens Schaefer, Şükrü Burç Eryilmaz, Wenqiang Zhang, Dehai Wu, Stephen Deiss, Priyanka Raina, He Qian, Bin Gao, Siddharth Joshi, Huaqiang Wu, Hon-Cheng Wong, and Gert Cauwenberghs. A compute-in-memory chip based on resistive random-access memory. *Nature*, 608(7923):504–512, 8 2022. doi: 10.1038/s41586-022-04992-8. URL `https://doi.org/10.1038/s41586-022-04992-8`.

Shihui Yin, Yulhwa Kim, Xu Han, Hugh Barnaby, Shimeng Yu, Yandong Luo, Wangxin He, Xiaoyu Sun, Jae-Joon Kim, and Jae-sun Seo. Monolithically integrated rram- and cmos-based in-memory computing optimizations for efficient deep learning. *IEEE Micro*, 39(6):54–63, 2019. doi: 10.1109/MM.2019.2943047.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer F. Syeda-Mahmood, Anne L. Martel, Lena Maier-Hein, João Manuel R. S. Tavares, Andrew P. Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi (eds.), *DLMIA-MICCAI*, volume 11045, pp. 3–11, 2018.

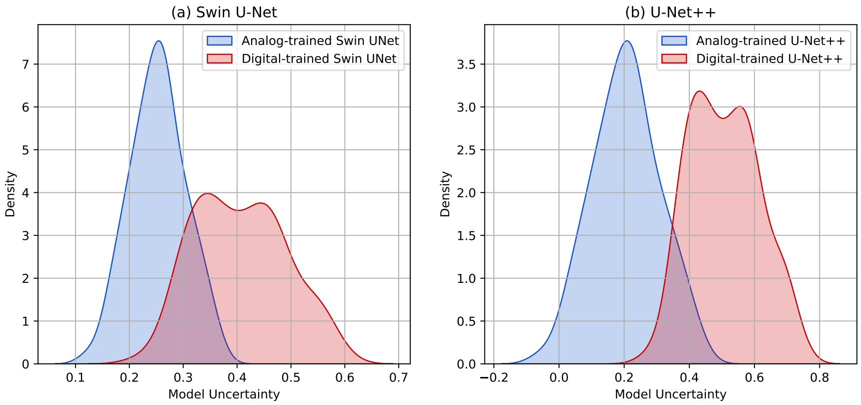


Figure 3: Model uncertainty density analysis across different brain tumor detection inputs.

# A Appendix

## A.1 Datasets

The Brain Tumor Segmentation dataset from The Cancer Imaging Archive (Simpson et al., 2019) features MRI scans and segmentation masks for 110 lower-grade glioma patients, enriched with FLAIR sequences and genomic cluster data. The Spleen Segmentation dataset, also from the Medical Segmentation Decathlon (Simpson et al., 2019), offers CT scans for detailed spleen segmentation, focusing on organ delineation. Lastly, the Nuclei detection dataset (Caicedo et al., 2019) provides a broad collection of segmented nuclei images, showcasing diversity in cell types and imaging modalities, including brightfield and fluorescence techniques.

## A.2 Models & training

U-Net, effective in biomedical segmentation, offers a balanced ’U’ shaped structure for localization. U-Net++ enhances U-Net by adding nested and skip pathways for improved detail capture. Swin Transformer, unlike these, utilizes shifted windows to handle image patches, focusing on long-range interdependencies. Transitioning from initial training, the models were adapted to hardware-aware training through noise injection, facilitated by AIHWKIT (Rasch et al., 2021). This approach, supported by MONAI (Cardoso et al., 2022) frameworks and other open-source tools, optimized our models for in-memory computing applications. The simulation of the analog aware training’s parameters and characteristics is available in the provided code and defined in the function $create\_rpu\_config()$.

## A.3 Extended Certainty analysis

Figure 3 provides an overall certainty analysis of U-Net++ and Swin U-Net when trained with hardware training versus digital training. While the main paper presents a compelling and illustrative example, we extend the density quantification to the whole test set. The uncertainty is computed using Monte Carlo Sampling. This involves using multiple passes of the input data and observing the variability in the outputs. The lower the variability, the higher the certainty of the model’s predictions.

Results suggest that the analog-trained model exhibits a lower uncertainty density compared to the digital-trained model. This implies that the analog-trained U-Net++ is be more reliable and consistent in its predictions, making it a more suitable choice for medical applications. This is mainly due to the hardware-aware training that forces the model to adapt to the inherent variability and constraints of the analog computing environment, thereby enhancing its ability to handle uncertain scenarios with greater precision.