This the accepted version of a book chapter. Please cite it as follows:
Busia, P., Pinna, A., Meloni, P. (2025). Endoscopy Image Classification for Wireless Capsules with CNNs on Microcontroller-Based Platforms. In: Lorandel, J., Kamaleldin, A. (eds) Design and Architecture for Signal and Image Processing. DASIP 2025. Lecture Notes in Computer Science, vol 15569. Springer, Cham.
https://doi.org/10.1007/978-3-031-87897-8_5

Endoscopy Image Classification for Wireless Capsules with CNNs on Microcontroller-based Platforms

Paola Busia[1 [0000−0002−1434−9858]], Andrea Pinna[2 [0000−0001−5369−0787]], and Paolo Meloni[1 [0000−0002−8106−4641]]

[1] DIEE, University of Cagliari, 09123 Cagliari, Italy
[2] Sorbonne University, CNRS, LIP6, 75005 Paris, France
paola.busia@unica.it, andrea.pinna@lip6.fr, paolo.meloni@unica.it

Abstract. Wireless Capsule Endoscopy (WCE) offers an important diagnostic instrument for different gastrointestinal diseases. Enhancing the WCE device with real-time image processing capabilities allows to assist specialized physicians in the long and cumbersome process of inspecting the significant amount of data acquired during the examination procedure, providing the first detection of the signs of relevant diseases that require further attention. In this work, we evaluate different state-of-the-art Convolutional Neural Network models for real-time WCE image classification, focusing on lightweight topologies suitable for execution on low-power microcontroller platforms and integration on the WCE device. The selected WCE-SqueezeNet model achieves 98.5% accuracy in the classification of ulcerative colitis, polyps, and esophagitis against healthy samples, allowing classification at a 16 fps rate on the GAP9 multi-core platform, with 61 ms inference time and 30.6 mW average core power consumption.

Keywords: Wireless Capsule Endoscopy · Near-Sensor Processing · Convolutional Neural Networks

1 Introduction

Gastrointestinal (GI) diseases represent a relevant concern for the health of millions of people worldwide. As a reference, the number of patients living with a GI condition in Europe between 2000 and 2019 was estimated to be over 332 millions [1]. Wireless Capsule Endoscopy (WCE) represents a common diagnostic instrument for the early detection of GI diseases, which enables proper medical intervention before more serious complications arise.

The current examination procedure involves image acquisition with the WCE device along the GI tract, allowing the detailed exploration of the tissue, and the transmission of a huge amount of image data to an external server for direct medical examination. Image processing and classification of common conditions through Artificial Intelligence (AI) approaches have been evaluated in the literature, with the aim of assisting the physicians in the diagnostic process [8, 10, 2].

Nonetheless, the continuous stream of images to the server through the wireless channel requires a significant amount of bandwidth, thus in-place processing has been explored [11], based on the general AI-at-the-edge trend in medical applications [3].

This work focuses on enabling near-sensor image processing capabilities on the WCE device, in order to perform real-time image classification and limit the WCE-to-server transmission only to the images representing a recognized symptom of the disease. To this aim, we target image classification based on Convolutional Neural Networks (CNNs) with a complexity suitable for efficient inference execution on low-power microcontroller-based platforms, that are compatible with the integration on the WCE device. Recent AI-oriented platforms in the edge domain leverage up to a few MBs of available memory [13], although the working memory of most common microcontroller-based platforms is within 1 MB. The targeted computing platforms thus introduce a significant constraint on the complexity of the classification model, both in terms of memory requirements and the number of required operations to ensure real-time processing.

The contributions of this work can thus be summarized in two main points:

- the evaluation on an open-source dataset of a suitable CNN classifier, the WCE-SqueezeNet model, for real-time near-sensor WCE image classification, reaching 98.5% accuracy in the recognition of three common GI conditions, including the assessment of the most effective image resolution reduction as a trade-off between accuracy and computational complexity;
- the preparation, deployment, and demonstration of real-time execution on the GAP9 low-power multi-core platform, enabling a 16 fps throughput within a core power envelope of 30.6 mW.

The paper is organized as follows: Section 2 summarizes the state of the art, Section 3 describes our proposed approach for the classification model training and evaluation, Section 4 reports the experimental results, and finally Section 5 summarizes the conclusions.

2 Related Work

The use of AI, particularly of CNN models, for medical image processing is well documented in the literature. Table 1 summarizes recent works from the state of the art, addressing WCE-image classification with CNN models, and referencing the same set of open source data, with the exception of [11].

The author of [8] presents a classification model obtained from the combination of truncated versions of the EfficientNetB0, MobileNetV2, and ResNet50V2 topologies, exploited as feature extractors prior to a Fusion Residual Block producing classification. The classification model, called MFuRE-CNN, reaches 97.75% accuracy in the recognition of 3 pathological conditions against healthy samples.

The EfficientNet topology is also exploited in the work of [10], where the EfficientNetV2B2 topology is adapted to the WCE task with the integration of a custom classification head including Global Average Pooling and Dense layers. The

finally obtained GastroNet model outperforms the evaluated alternatives fine-tuned from state-of-the-art topologies, such as ResNet50, and EfficientNetv2B1, reaching over 99% accuracy.

Additionally, the work of [2] presents the DCDS-Net model, exploiting several blocks of separable convolutions prior to a classification block composed of three Dense layers.

As can be noticed from the table, the efficiency of the classifiers described in these works was not assessed on an embedded hardware target, however, due to their storage requirements, of at least 20 MB, and the number of operations required per inference run, over 2 GOPS, they do not represent suitable candidates for integration in an intelligent WCE device. On the other hand, the idea of exploiting network models reaching state-of-the-art accuracy on the ImageNet dataset was demonstrated as a successful approach, to be considered also for WCE classification.

A system based on real-time image processing on the WCE device is envisioned and assessed in the work of [11], where memory constraints and efficiency are taken into account for the selection of the CNN detection model, showing 99.5% average precision in the recognition and detection of colorectal polyps. The precision number refers to a 25% intersection-over-union between the detected bounding box and the ground truth, evaluated on the data acquired from 255 patients of Denmark's national screening program. The number of parameters is still significantly high, over 3 million, but it is compatible with the proposed camera-pill hardware architecture, integrating 8 MB of memory, and where the average power consumption was assessed to be around 50 mW.

In this work, we target a similar problem, aiming at real-time classification of different GI diseases. Exploiting the feature extraction capabilities, derived from learning on large image datasets, of the pre-trained state-of-the-art SqueezeNet network, our proposed WCE-SqueezeNet classification model reaches a competitive accuracy compared to the alternatives, within a complexity and memory footprint suitable for inference deployment on resource-constrained low-power hardware platforms, demonstrated through direct measurements. The efficiency of the proposed model surpasses the topology presented in [11], both in terms of required parameters and operations.

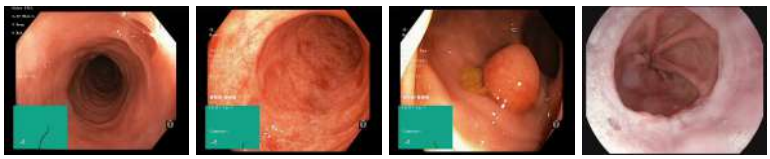
Table 1: Comparison with the state of the art of WCE image classification models.

Model	Accuracy	Precision	Recall	Deployed	Parameters Memory	GOPS*
MFuRE-CNN [8]	97.75%	97.75%	97.75%	✗	19.2 MB	7.8
ResNet50 [10]	98%	98.1%	98%	✗	89 MB**	7.6**
EfficientNetV2B1 [10]	98.5%	98.5%	98.5%	✗	25.9 MB**	2**
GastroNetV1 [10]	99.2%	99.3%	99.3%	✗	32.8 MB**	2.3**
DCDS-Net [2]	99.33%	99.37%	99.32%	✗	83.8 MB	9**
YOLO-based [11]	99.5 AP	N.R.	N.R.	✓	3.2 MB	0.9**
this work	98.5%	98.6%	98.5%	✓	750 kB	0.45

* The number of operations refers separately to multiplications and additions: 1 MAC = 2 OPS.
** Estimated from paper.

3 Method

This section describes our training and evaluation approach for the development of the proposed classification system, presenting the reference dataset and target hardware platform, as well as the CNN model considered for the assessment.



(a) (b) (c) (d)

Fig. 1: Sample images from the reference dataset [9, 14], including a) normal sample, b) ulcerative colitis, c) polyp, d) esophagitis.

3.1 Dataset

This study references the KVASIR [9] and the ETIS-Larib Polyp [14] databases, according to the data organization introduced in [8]. This collection counts 6000 images acquired through WCE, including an equal number of examples for three main pathological conditions of the gastrointestinal tract, such as ulcerative colitis, polyps, and esophagitis, as well as healthy/normal samples. Figure 1 reports an example of acquired image for each of the targeted classes.
Ref. [8] also introduced a standard training, validation, and test split, according to the scheme summarized in Table 2. All the images in the dataset were pre-processed in order to standardize their size to a 224×224 resolution and normalized according to the data format expected by the different models considered. The selection of the input resolution was then the subject of a dedicated exploration, which is described in detail in Section 4.

Table 2: Data organization into training, validation, and test set.

Class	Train	Valid	Test
Normal - N	800	500	200
Ulcer - U	800	500	200
Polyps - P	800	500	200
Esophagitis - E	800	500	200
Tot	**3200**	**2000**	**800**

3.2 Hardware Target

To evaluate the efficiency of the WCE image classification application, we consider as a target the GAP9 Parallel Ultra-Low-Power (PULP) platform [4]. This

device recently demonstrated remarkable energy efficiency in the tiny-ML benchmarks [7], with 0.033mW/GOP. It is an advanced microcontroller-based platform, integrating a cluster of nine parallel processors, which have access to a shared 128 kB L1 memory. The cluster can be exploited for parallel processing and the acceleration of typical deep learning workloads, such as convolutional and fully connected layers. The memory hierarchy also includes a 1.5 MB L2 memory, thus providing enough storage and computational resources to accommodate the classification model, within the limited power budget compatible with integration on WCE devices.

3.3 Classification Approach

A common approach in medical image classification problems is to leverage the feature extraction capabilities of off-the-shelf models pre-trained on large image datasets, such as ImageNet, and finally specialize them for the task at hand [6]. This solution often results in higher classification performance than training the same topology from scratch, as the available medical data is typically reduced and sometimes unbalanced in the representation of the different conditions.

Considering these documented results, in this work we aim to fine-tune an image classification model for WCE image classification. The network topology was selected based on the assessment of the computational and storage requirements, with the aim of targeting ultra-low-power deployment on tiny microcontroller-based platforms, to perform near-sensor image processing directly on the WCE device. We thus defined a memory constraint of 1 MB as the maximum acceptable memory footprint, limiting the evaluation to network models exploiting less than 1 million parameters. Therefore, we selected the SqueezeNet topology [5] as the backbone of our classification model. The structure of the model is recalled in Figure 2. Compared to the Vanilla topology, we replaced the classification head with a dense layer suitable for the new 4-classes problem. In this configuration, the model exploits less than 740k parameters, thus allowing to meet the memory constraint with 8-bit quantization.

For the fine-tuning of the model, we leveraged the Pytorch framework, exploiting CrossEntropy loss, SGD optimizer, and 0.001 learning rate. The learning rate was iteratively adapted after patience of 30 epochs without improvements on the validation set, considering up to 6 steps. The final update was considered as the early stop condition. The training was performed on Google Colaboratory, leveraging the T4 GPU.

4 Experimental Results

In this section, we summarize the experimental results obtained on the target dataset, highlighting the most relevant metrics describing the classification performance, and including the impact of quantization on the final accuracy. We finally assess the efficiency of the selected solution with on-hardware direct measurements, demonstrating the feasibility of performing real-time classification on suitable low-power hardware targets.

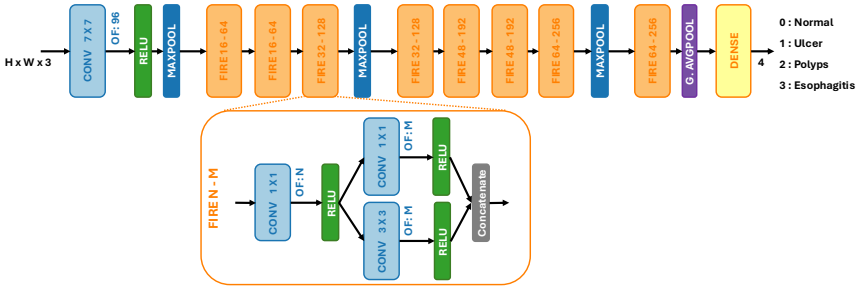


Fig. 2: WCE-SqueezeNet classification model.

4.1 Classification Performance Assessment

As the first step of the performance investigation, we compared the achievable accuracy when training and testing the classification model on images of different resolutions, starting from the 224×224 resolution exploited in [8], then reducing the size of the training image to 128×128, and finally to 64×64. The outcome of the exploration is summarized in Figure 3, where the evaluated alternatives are placed according to the accuracy achieved on the validation set and to their computational complexity in terms of number of required operations (GOPS). Each model in the plot was evaluated on images of the same resolutions as the examples learned during the training. As can be observed, reducing the image size to 128×128 introduces only a negligible drop in the accuracy, while resulting in a significant reduction, by a factor of 3×, of the computational workload. On the contrary, the performance degrades significantly when the resolution is reduced further.

Based on this first assessment, and considering the low-power hardware targeted for the deployment, we selected the model trained to perform classification on the 128×128 images. We then performed a refinement training, based on the exploration of the most relevant hyperparameters, with batch size equal to 128, reaching an accuracy of 99% for full precision inference on the test set.

The details of the confusion matrix obtained with the WCE-SqueezeNet model are reported in Table 3a. As can be observed, the model shows perfect specificity, providing 100% recall in the recognition of the normal condition, with no false alarms raised based on the examples in the test set. The average precision and recall on all the targeted classes are 99%. Due to the importance of polyps' early detection, we further explored the accuracy in the recognition of this target class. Figure 4 shows the Receiver Operating Curve for the recognition of the polyps class against all other classes in the dataset. As can be noticed, the model provides a good discrimination ability, with an area under the curve (AUC) value equal to 0.99.

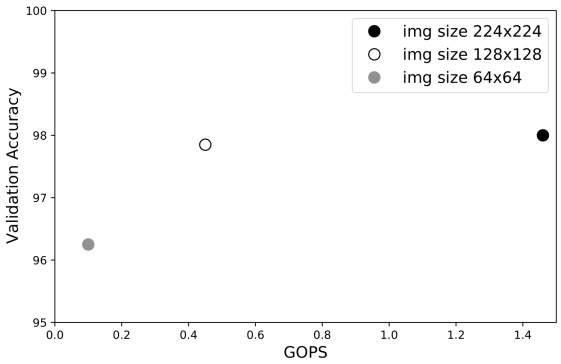


Fig. 3: Exploration of the input image resolution considering the validation set.

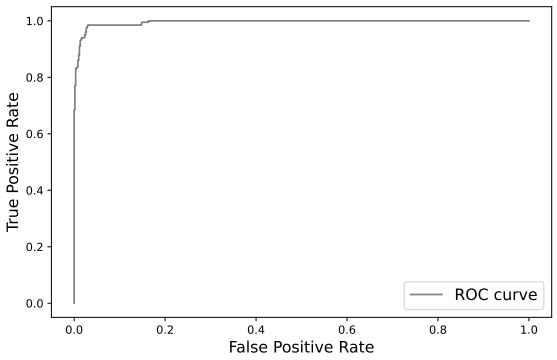


Fig. 4: Receiver Operating Characteristic curve for polyps recognition.

Table 3: Confusion Matrix resulting from test set classification with the WCE-Squeezenet model.

(a) Full Precision.

True Labels \ Predicted Labels	N	U	P	E
Normal	**200**	0	0	0
Ulcer	0	**198**	2	0
Polyps	1	4	**195**	0
Esophagitis	0	1	0	**199**

(b) 8-bit Integer.

True Labels \ Predicted Labels	N	U	P	E
Normal	**200**	0	0	0
Ulcer	0	**197**	3	0
Polpys	3	4	**193**	0
Esophagitis	0	1	1	**198**

4.2 Quantization

As anticipated, the aim of this work is to enable real-time classification on the WCE device, targeting execution on low-power microcontrollers. In order to meet the memory constraint of 1 MB, the memory requirements of the WCE-SqueezeNet model needed to be reduced through quantization to 8-bit precision. The quantization was performed through the TensorflowLite utilities, resulting in a limited accuracy drop, to 98.5%, compared to the Floating Point 32-bit full precision representation. The confusion matrix obtained on the test set with the integer model is reported in Table 3b. As can be noticed, only a few errors involving the pathological classes were introduced.

4.3 Discussion

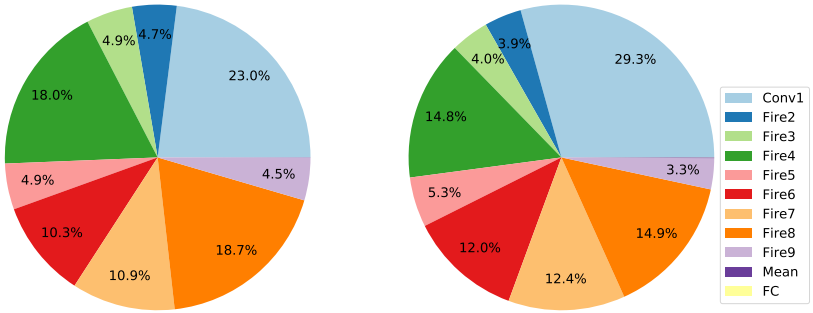
Table 4: Classification Performance Assessment on the test set. PT column indicates whether the training started from ImageNet trained weights.

Model	PT	Accuracy	Ulcer		Polyps		Esophagitis	
			Precision	Recall	Precision	Recall	Precision	Recall
WCE-Squeezenet	✓	**98.13**	98.45	95	95.12	97.5	100	100
WCE-Squeezenet	✗	84.63	64.84	88.5	81.97	50	98	100
Squeezenet + SVM	✓	96.5	95	95.5	95.29	91	100	100
WCE-MobileNetV2	✓	**98.5**	97	97	97	97	100	100
MobileNetV2 + SVM	✓	96.87	96.39	93.5	94	94	100	100

In this section, we discuss the effectiveness of the classification approach described in Section 3. First, we evaluate the training approach, comparing it to the classification performance achievable with:

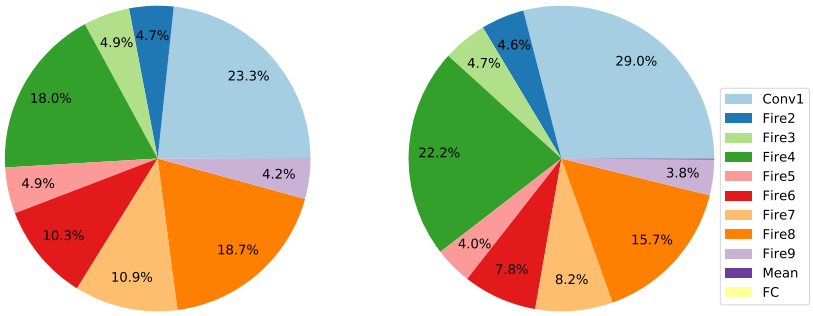
- the same topology trained from scratch, with no previous knowledge acquired on the ImageNet dataset;
- the pre-trained feature-extraction model, combined with a classifier trained on the target problem.

Additionally, we also considered the comparison with a more complex network model, such as MobileNetV2 [12]. Table 4 summarizes the comparison, based on models trained on images of 224×224 resolution. As can be observed, standalone CNN classification outperforms the classification of the extracted features based on SVM for both the network models considered. Furthermore, starting from the pre-trained parameters provides a significant advantage over training the same topology from scratch. Finally, the classification performance enabled by the WCE-SqueezeNet model is very close to the best one achievable with the more complex MobileNet topology.



(a) OPS. (b) Execution Cycles.

Fig. 5: Computational workload of the different layers in WCE-SqueezeNet model evaluated in terms of number of required operations and required execution cycles on the GAP9 platform, for 224×224 input resolution.



(a) OPS. (b) Execution Cycles.

Fig. 6: Computational workload of the different layers in WCE-SqueezeNet model evaluated in terms of number of required operations and required execution cycles on the GAP9 platform, for 128×128 input resolution.

4.4 Deployment

The deployment of the selected WCE-SqueezeNet model was automated through the proprietary code generation tool, the GAP9 SDK. We compared the required inference time for the model applied to 224×224 and to 128×128 input images, considering parallel execution on eight cores of the computing cluster. In the first case, inference time was measured equal to 0.2 s, thus resulting in an expected throughput of 5 fps, evaluated at a 370 MHz working frequency. The average computational efficiency was 9 OPS/cycle. In the second case, the reduced computational workload resulted in only 61 ms inference time, with an expected throughput of 16 fps and an average computational efficiency of 10 OPS/cycle.

The composition of the computational workload is reported in Figure 5a and 6a, whereas the required inference time on the GAP9 platform for each layer

is represented in Figure 5b and 6b. As can be noticed, the composition of the expected workload based on the number of required operations and of the measured inference time is very similar, showing there is no significant inefficiency in the implementation of the most relevant operands. The computational workload is dominated by the convolutional layers, as the main operands exploited in the Fire modules, while the contribution of the fully connected classification head is negligible. Input resolution shows only a limited impact on the composition of the required execution time.

Finally, we assessed the energy efficiency, by measuring the average core power consumption during inference execution, which is equal to 30.6 mW. The required energy per inference is thus 1.9 mJ. Measurements were performed with Nordic Power Profiler II. This result demonstrates the suitability of performing real-time inference on the WCE device, with limited power requirements compatible with battery-powered solutions.

5 Conclusions

In this work, we presented a classification model for the recognition of GI diseases based on WCE acquisition. The WCE-SqueezeNet model demonstrated 98.5% classification accuracy, evaluated after 8-bit quantization for efficient inference execution on the targeted GAP9 low-power platform. The analysis of the accuracy degradation with the progressive reduction of the input image resolution showed that compression up to a 128×128 resolution is possible with a negligible impact on the accuracy. The efficiency of the proposed solution was evaluated on the GAP9 platform, considering parallel execution on 8 cores. The measurements demonstrated a 16 fps achievable throughput and an average core power consumption of 30.6 mW, compatible with possible integration in the WCE device.

Acknowledgments. We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No.3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR). This research was also supported by the French National Research Agency (ANR) under the LabCom program 2021 - V2 (ICI-Lab) and by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement GA 101140052 (H2TRAIN).

References

1. Tackling the burden of digestive disorders in europe. The Lancet Gastroenterology & Hepatology **8**, 95 (2023). https://doi.org/10.1016/S2468-1253(22)00431-9, https://doi.org/10.1016/S2468-1253(22)00431-9

2. Asif, S., Zhao, M., Tang, F., Zhu, Y.: Dcds-net: Deep transfer network based on depth-wise separable convolution with residual connection for diagnosing gastrointestinal diseases. Biomedical Signal Processing and Control **90**, 105866 (2024). https://doi.org/https://doi.org/10.1016/j.bspc.2023.105866, https://www.sciencedirect.com/science/article/pii/S1746809423012995
3. Busia, P., Scrugli, M.A., Jung, V.J.B., Benini, L., Meloni, P.: A tiny transformer for low-power arrhythmia classification on microcontrollers. IEEE Transactions on Biomedical Circuits and Systems pp. 1–11 (2024). https://doi.org/10.1109/TBCAS.2024.3401858
4. Greenwaves: Ultra low power gap processors (October 2024), https://greenwaves-technologies.com/low-power-processor/
5. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. arXiv:1602.07360 (2016)
6. Li, X., Cen, M., Xu, J., Zhang, H., Xu, X.S.: Improving feature extraction from histopathological images through a fine-tuning imagenet model. Journal of Pathology Informatics **13**, 100115 (2022). https://doi.org/https://doi.org/10.1016/j.jpi.2022.100115, https://www.sciencedirect.com/science/article/pii/S215335392200709X
7. ML Commons: Inference: tiny. v1.0 Results (2024), https://mlcommons.org/en/inference-tiny-10/, Accessed: 30-10-2024
8. Montalbo, F.J.P.: Diagnosing gastrointestinal diseases from endoscopy images through a multi-fused cnn with auxiliary layers, alpha dropouts, and a fusion residual block. Biomedical Signal Processing and Control **76**, 103683 (2022). https://doi.org/https://doi.org/10.1016/j.bspc.2022.103683, https://www.sciencedirect.com/science/article/pii/S1746809422002051
9. Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.T., Lux, M., Schmidt, P.T., Riegler, M., Halvorsen, P.: Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference. p. 164–169. MMSys'17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3083187.3083212, https://doi.org/10.1145/3083187.3083212
10. Rajkumar, S., Harini, C.S., Giri, J., Sairam, V.A., Ahmad, N., Badawy, A.S., Krithika, G.K., Dhanusha, P., Chandrasekar, G.E., Sapthagirivasan, V.: GastroNet: A CNN based system for detection of abnormalities in gastrointestinal tract from wireless capsule endoscopy images. AIP Advances **14**(8), 085223 (08 2024). https://doi.org/10.1063/5.0208691, https://doi.org/10.1063/5.0208691
11. Sahafi, A., Wang, Y., Rasmussen, C.L.M., Bollen, P., Baatrup, G., Blanes-Vidal, V., Herp, J., Nadimi, E.S.: Edge artificial intelligence wireless video capsule endoscopy. Scientific Reports **12**, 13723 (2022). https://doi.org/10.1038/s41598-022-17502-7, https://doi.org/10.1038/s41598-022-17502-7
12. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
13. Scherer, M., Macan, L., Jung, V.J.B., Wiese, P., Bompani, L., Burrello, A., Conti, F., Benini, L.: Deeploy: Enabling energy-efficient deployment of small language models on heterogeneous microcontrollers. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **43**(11), 4009–4020 (2024). https://doi.org/10.1109/TCAD.2024.3443718

14. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. vol. 9, pp. 283 – 293 (2014). https://doi.org/10.1007/s11548-013-0926-3, https://doi.org/10.1007/s11548-013-0926-3