

MedSegBench: A Comprehensive Benchmark for Medical Image Segmentation in Diverse Data Modalities

Musa Aydin^{1,†} and Zeki Kuş^{1,†}

¹Fatih Sultan Mehmet Vakif University, Computer Engineering, İstanbul, 34445, Türkiye

*corresponding author(s): Zeki Kuş (zkus@fsm.edu.tr)

[†]these authors contributed equally to this work

ABSTRACT

MedSegBench is a comprehensive benchmark designed to evaluate deep learning models for medical image segmentation across a wide range of modalities. This benchmark includes 35 datasets with over 60,000 images, covering modalities such as ultrasound, MRI, and X-ray. It addresses challenges in medical imaging, such as variability in image quality and dataset imbalances, by providing standardized datasets with train/validation/test splits. The benchmark supports binary and multi-class segmentation tasks with up to 19 classes. Evaluations are conducted using the U-Net architecture with various encoder/decoder networks, including ResNets, EfficientNet, and DenseNet, to evaluate model performance. MedSegBench serves as a valuable resource for developing robust and flexible segmentation algorithms. It allows for fair comparisons across different models and promotes the development of universal models for medical tasks. The datasets and source code are publicly available, encouraging further research and development in medical image analysis.

Background & Summary

Deep learning has become essential in medical image analysis and segmentation, offering powerful methods to help doctors and researchers better understand and diagnose diseases¹. Deep learning can identify patterns and details in medical images that might be difficult for human eyes to detect using complex networks such as convolutional neural networks². These techniques are precious for finding tumors in X-rays, classifying different cell types in whole-slide images, or segmenting different brain parts in MRI scans. However, working with biomedical datasets presents unique challenges, including variability of image quality and resolution, the need for well-annotated examples, imbalances of the datasets, and different modalities. Addressing these challenges and ensuring the effectiveness of deep learning methods in real-world medical settings requires large and diverse datasets³. These comprehensive collections of medical images help train the algorithms to handle different modalities and medical tasks. They also allow researchers to compare deep learning methods fairly, determine the most effective approaches for specific medical tasks, and develop universal models for different medical tasks.

There are limited benchmark studies in the literature focused on medical imaging, with most concentrating on medical image classification problems⁴⁻⁸. Gelasca et al.⁴ present a comprehensive biomedical segmentation benchmark that evaluates bioimage analysis methods. It includes six datasets with associated ground truth and validation methods, covering different scales from subcellular to tissue levels. Rebuffi et al.⁵ propose the Visual Decathlon Challenge, a benchmark that evaluates models across ten diverse visual classification domains, including datasets such as Aircraft, CIFAR-100, and ImageNet. Medical Segmentation Decathlon⁶ supports the creation and benchmarking of semantic segmentation algorithms. It includes 2633 3D images from ten different anatomical sites and modalities collected from multiple institutions and annotated by experts. Yang et al.⁷ introduce the MedMNIST Benchmark, a collection of ten pre-processed medical image datasets standardized to 28×28 pixels. It covers various medical image modalities and support multiple classification tasks. Yang et al.⁸ extend MedMNIST with MedMNIST v2, a standardized collection of biomedical image datasets. This includes 12 datasets for 2D images and 6 for 3D images, covering various data modalities, scales, and classification tasks.

This study introduces a comprehensive benchmark dataset for medical image segmentation (Figure 1). It includes 35 distinct datasets with over 60,000 images covering various data modalities such as ultrasound, dermoscopy, MRI, X-ray, OCT, and more. It provides a diverse resource for evaluating the performance of deep learning models in medical image segmentation tasks. The dataset includes a wide range of scales, from small collections with just a few dozen images to extensive datasets containing tens of thousands of samples. The segmentation tasks cover both binary and multi-class problems, with some datasets featuring up to 19 different classes. This benchmark offers several powerful advantages as a robust and versatile tool

38 for the research community:

- 39 • **Diversity of modalities:** The benchmark includes datasets from various imaging modalities such as Ultrasound, MRI,
40 X-Ray, OCT, Dermoscopy, Endoscopy, and various types of microscopy.
- 41 • **Task complexity:** It covers both binary segmentation tasks and multi-class segmentation tasks with up to 19 classes.
- 42 • **Dataset sizes:** There's a wide range in the number of images per dataset, from as few as 28 to as many as 21,165.
- 43 • **Data split:** All datasets follow a standard train/validation/test split, which is crucial for the proper evaluation of machine
44 learning models.
- 45 • **Standardization:** All datasets are standardized to enhance comparability and ease of use. Samples across all datasets
46 have been resized to three standard resolutions - 128, 256, and 512 pixels - and stored in a uniform format.
- 47 • **Application areas:** The datasets cover various medical applications, including cancer detection, COVID-19 diagnosis,
48 cell and nuclei segmentation, and organ segmentation.

49 We have evaluated each dataset on state-of-the-art segmentation model (U-Net⁹) with different encoder/decoder network
50 types (ResNet-18, ResNet-50, Efficient-Net, MobileNet-v2, DenseNet-121, Mix Vision Transformer)¹⁰. Each experiment are
51 performed 3 times and average results are reported.

52 This benchmark is carefully designed to assess how well deep learning models can generalize across different medical
53 domains, perform on small and large datasets, and handle varying task complexities. By including such a wide array of medical
54 imaging challenges, this benchmark is a powerful tool for comprehensively evaluating the robustness, flexibility, and overall
55 efficacy of segmentation algorithms in the medical imaging field.

56 Methods

57 Data Preparation

58 The MedSegBench dataset comprises 35 distinct 2D medical image segmentation datasets, some of which are extracted from
59 3D slices. These datasets cover various data modalities such as Ultrasound, OCT, Chest X-ray, MR, and more. The original
60 datasets differ in scales, segmentation tasks (binary/multi-class), classes, imaging modalities, and annotation styles. Hence, we
61 have selected a standardized format and performed pre-processing to ensure a consistent format across all datasets.

62 Numerous medical image segmentation datasets are available in the literature, each presenting various challenges in-
63 cluding variations in annotations, image sizes, and file formats. Additionally, many of these datasets lack officially shared
64 train/test/validation splits, making it challenging to fairly compare different methods. To address these issues, we performed
65 pre-processing steps. All image and label pairs are resized to 128×128 , 256×256 , and 512×512 pixels using the bicubic
66 interpolation method. Although we used 512×512 sized images in our experiments, we have made the 128×128 and 256×256
67 sized versions publicly available for researchers with limited GPU memory. Also, we have applied a mapping to labels; pixels
68 with values of 0 and 255 are mapped to 0 and 1 for binary segmentation tasks, and for multi-class segmentation tasks, pixels are
69 mapped to integer values between 0 and (#Classes - 1). No additional augmentation or pre-processing steps are applied to the
70 images and labels. We have followed three different scenarios based on MedMNIST v2⁸ to create train/test/validation splits: (1)
71 Utilizing the source train/test/validation splits if published by the authors; (2) Using the source validation set as the test set and
72 splitting the source training set into 90% training and 10% validation (9:1 ratio) if the source training and validation splits are
73 published by the authors; (3) Randomly splitting the dataset into 70% training, 10% validation, and 20% test sets if no public
74 train/test/validation splits are available (7:1:2 ratio). Most of these datasets are publicly published under Creative Commons
75 Licenses, some of which are CC-BY-NC, CC-BY-SA, and CC-BY-NC-SA, permitting the redistribution of datasets. We have
76 published datasets in MedSegBench under Creative Commons Licences, and source codes have been published under Apache
77 License 2.0.

78 Table 1 presents the summary information for all MedSegBench datasets. In addition, Table 2 shows the data-modality-
79 based overview for MedSegBench datasets. Furthermore, Table 3 provides an overview of various datasets, detailing their
80 sub-categories and the number of samples for training, validation, and testing. In the following sections, we will describe the
81 details of each dataset.

82 Details

83 **AbdomenUSMSBench:** The AbdomenUSMSBench created from AbdomenUS^{11,12} consists of 926 ultrasound images of
84 the abdominal region, each with a resolution of 449×464 pixels. This dataset is designed for multi-class segmentation tasks
85 and includes eight distinct classes. We have used the official train and test splits, and the train set is split into a training and

Table 1. Overview of the MedSegBench datasets, including source references, modality, task types (binary or multi-class) with number of classes, total sample sizes and train/validation/test splits.

Dataset Name ^{source}	Modality	Binary or Multi-class (# Classes)	# Images	# Train/Val/Test
AbdomenUSMSBench ^{11,12}	Ultrasound	Multi-class (8)	926	569/64/293
Bbbc010MSBench ^{13,14}	Microscopy	Binary	100	70/10/20
Bkai-Igh-MSBench ¹⁵⁻¹⁷	Endoscopy	Multi-class (3)	1,000	700/100/200
BriFiSegMSBench ^{18,19}	Microscopy	Binary	1,360	1005/115/240
BusiMSBench ^{20,21}	Ultrasound	Binary	647	452/64/131
CellNucleiMSBench ^{22,23}	Nuclei	Binary	670	469/67/134
ChaseDB1MSBench ²⁴	Fundus	Binary	28	19/2/7
ChuacMSBench ²⁵	Fundus	Binary	30	21/3/6
Covid19RadioMSBench ²⁶⁻²⁸	Chest X-Ray	Binary	21,165	14,814/2,115/4,236
CovidQUEXMSBench ^{29,30}	Chest X-Ray	Binary	2,913	1,864/466/583
CystoFluidMSBench ³¹⁻³³	OCT	Binary	1,006	703/101/202
Dca1MSBench ^{34,35}	Fundus	Binary	134	93/13/28
DeepbacsMSBench ^{36,37}	Microscopy	Binary	34	17/2/15
DriveMSBench ^{38,39}	Fundus	Binary	40	18/2/20
DynamicNuclearMSBench ^{40,41}	Nuclear Cell	Binary	7,084	4,950/1,417/717
FHPsAOPMSBench ^{42,43}	Ultrasound	Multi-class (3)	4,000	2,800/400/800
IdribMSBench ^{44,45}	Fundus	Binary	80	47/6/27
Isic2016MSBench ^{46,47}	Dermoscopy	Binary	1,279	810/90/379
Isic2018MSBench ⁴⁸⁻⁵⁰	Dermoscopy	Binary	3,694	2,594/100/1,000
KvasirMSBench ^{51,52}	Endoscopy	Binary	1,000	700/100/200
M2caiSegMSBench ^{53,54}	Endoscopy	Multi-class (19)	614	245/307/62
MonusacMSBench ^{55,56}	Pathology	Multi-class (6)	310	188/21/101
MosMedPlusMSBench ^{57,58}	CT	Binary	2,729	1,910/272/547
NucleiMSBench ⁵⁹	Pathology	Binary	141	98/14/29
NusetMSBench ^{60,61}	Nuclear Cell	Binary	3,408	2,385/340/683
PandentalMSBench ^{62,63}	X-Ray	Binary	116	81/11/24
PolypGenMSBench ^{64,65}	Endoscopy	Binary	1,412	984/140/288
Promise12MSBench ^{66,67}	MRI	Binary	1,473	1,031/147/295
RoboToolMSBench ³¹	Endoscopy	Binary	500	350/50/100
TnbnucleiMSBench ^{68,69}	Pathology	Binary	50	35/5/10
UltrasoundNerveMSBench ⁷⁰	Ultrasound	Binary	2,323	1,651/223/449
USforKidneyMSBench ^{71,72}	Ultrasound	Binary	4,586	3,210/458/918
UWSkinCancerMSBench ⁷³	Dermoscopy	Binary	206	143/19/44
WbcMSBench ^{74,75}	Microscopy	Multi-class (3)	400	280/40/80
YeazMSBench ^{76,77}	Microscopy	Binary	707	360/96/251

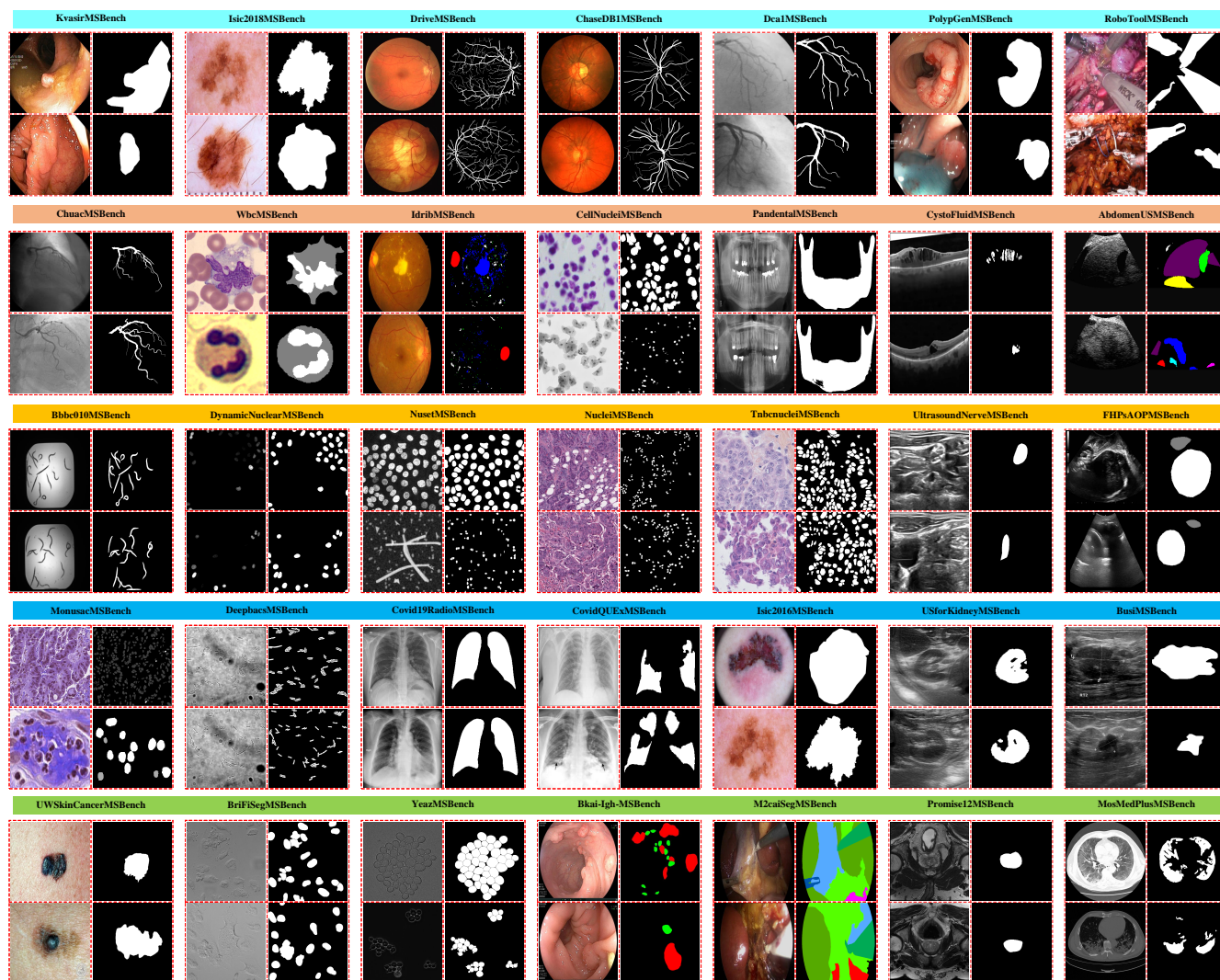


Figure 1. Summary of the MedSegBench

Table 2. Medical imaging modality and corresponding image counts

Modality	Number of Images
Computed Tomography	2,729
Dermoscopy	5,179
Endoscopy	4,526
Fundus	312
Magnetic Resonance Imaging	1,473
Microscopy	2,281
Nuclear Cell	10,492
Nuclei	670
Optical Coherence Tomography	1,006
Pathology	501
Ultrasound	12,482
X-Ray	24,194

86 validation set with a ratio of 9:1. The samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values
 87 between 0 and $(\#Classes - 1)$.

Table 3. Overview of datasets and their sub-categories with Train/Validation/Test splits. Each dataset is split into specific sub-categories by authors, and the corresponding number of samples for each sub-category is listed in Train/Val/Test format.

Dataset Name	Sub-categories	# Train/Val/Test
BriFiSegMSBench	C1: Target 1 A549; C2: Target 2 A549; C3: HeLa; C4: MCF7; C5: RPE1	201/23/48
BusiMSBench	C1: Benign; C2: Malignant	305/43/89 147/21/42
Covid19RadioMSBench	C1: Covid; C2: Lung; C3: Normal; C4: Viral Pneumonia	2,531/361/724 4,208/601/1,203 7,134/1,019/2,039 941/134/270
IdribMSBench	C1: Microaneurysms; C2: Hemorrhages; C3: Hard Exudates; C4: Optic Disc	47/6/27
UWSkinCancerMSBench	C1: Melenoma; C2: Not-Melenoma	83/11/25 60/8/19
WbcMSBench	C1: Lymphocyte; C2: Monocyte; C3: Neutrophil; C4: Eosinophil	146/20/43 63/9/43 44/6/13 23/3/8

Bbbc010MSBench: The Bbbc010MSBench dataset derived from Bbbc010^{13,14}, contains 100 microscopy images, each with a resolution of 696×520 pixels. These images are created for binary segmentation tasks and are originally captured for a screen in Fred Ausubel's Massachusetts General Hospital (MGH) lab. The dataset is split into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to 0 and 1.

Bkai-Igh-MSBench: The Bkai-Igh-MSBench dataset is derived from the BKAI-IGH NeoPolyp dataset¹⁵⁻¹⁷ and consists of 1,200 endoscopy images, each with a resolution of 1280×995 pixels. It is designed for multi-class segmentation tasks, with three distinct classes. We can not use publicly shared test sets because of a lack of ground truth annotations. The dataset is split into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and (#Classes - 1).

BriFiSegMSBench: The BriFiSegMSBench, which originates from the BriFiSeg dataset^{18,19}, includes 1,360 microscopy images with a resolution of 512×512 pixels. This dataset is intended for binary segmentation tasks and contains two classes. The images are single-channel samples derived from various cell lines, such as A549, HeLa, MCF7, and RPE1. The dataset is divided into training and validation sets with a 9:1 ratio. Additionally, task-specific images and annotations are provided in npz file format (see Table 3). The samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

BusiMSBench: The BusiMSBench dataset is derived from the Breast Ultrasound Images Dataset^{20,21} and contains 647 ultrasound images with an average resolution of 500×500 pixels. This dataset is designed for binary segmentation tasks, categorizing data into two classes: benign and malignant. It is split into three parts: train/val/test, in a 7:1:2 ratio. Additionally, class-based images (benign and malignant) and annotations are provided in .npz file format (see Table 3). The samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

CellNucleiMSBench: The CellNucleiMSBench comes from the 2018 Data Science Bowl^{22,23} and consists of 670 nuclei images with a resolution of 320×256 pixels. This dataset is specifically designed for binary segmentation tasks. We could not use 65 test images because ground truths are not published officially. Therefore, the source dataset split into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

ChaseDB1MSBench: ChaseDB1MSBench is based on the CHASE_DB1 dataset²⁴, released in 2012 by Kingston University, London, and St. George's, University of London, consists of 28 fundus images with a resolution of 999×960 pixels. This dataset is designed for binary segmentation tasks, including two classes. We split the source dataset into three parts:

116 train/val/test, in a 7:1:2 ratio. The samples are resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values
117 between 0 and 1.

118 **ChuacMSBench:** The ChuacMSBench, derived from the CHUAC dataset²⁵, includes 28 fundus images with 189×189
119 pixels. It is designed for binary segmentation tasks. The source dataset is split into three parts: train/val/test, in a 7:1:2 ratio.
120 The samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

121 **Covid19RadioMSBench** The COVID-19 Radiography Database²⁶⁻²⁸ is the source of the Covid19RadioMSBench dataset,
122 which consists of 21,165 chest X-ray images, each with a resolution of 299×299 pixels. This dataset is designed for binary
123 segmentation tasks. We divide the source dataset into three parts: train/val/test sets with a ratio of 7:1:2. It is developed by a
124 collaborative effort of researchers from Qatar University, the University of Dhaka, and partners from Pakistan and Malaysia,
125 working alongside medical professionals. It includes chest X-ray images for COVID-19 positive cases and Normal and Viral
126 Pneumonia images. The authors have also categorized the images into four groups: COVID, Lung_Opacity, Normal, and Viral
127 Pneumonia. We provide these category-based images and their corresponding annotations in .npz file format (see Table 3). The
128 samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

129 **CovidQUExMSBench:** The CovidQUExMSBench, based on the COVID-QU-Ex Dataset^{29,30}, consists of 2,913 chest
130 X-ray images, each with a resolution of 256×256 pixels. This dataset is specifically designed for binary segmentation tasks.
131 We use only infection segmentation samples. The source dataset is split into three parts: train/val/test, in a 7:1:2 ratio. The
132 samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

133 **MosMedPlusMSBench:** The MosMedPlusMSBench, based on the MosMedDataPlus^{57,58} dataset, comprises 2,729 Covid-
134 19 CT images, each sized 512×512 pixels. This dataset is designed for binary segmentation tasks. We split source data into
135 three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to integer
136 values between 0 and 1.

137 **CystoFluidMSBench:** The CystoFluidMSBench is based on Intraretinal Cystoid Fluid dataset³¹⁻³³, comprises 1,006
138 OCT (Optical Coherence Tomography) images, most of which are sized at 512×512 pixels. This dataset is designed for
139 binary segmentation tasks. The images are carefully chosen by medical experts at Liaquat University of Medical and Health
140 Sciences (LUMHS) Jamshoro, who are trained to identify Cystoid Macular Edema (CME) and its progression, providing a
141 confirmatory diagnosis of CME. The source dataset is split into three parts: train/val/test, in a 7:1:2 ratio. The samples are
142 resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

143 **Dca1MSBench:** The Dca1MSBench is derived from the DCA1 dataset^{34,35} and contains 134 fundus images, each with
144 a resolution of 300×300 pixels. The images are provided by the Cardiology Department of the Mexican Social Security
145 Institute, UMAE T1-León. This dataset is specifically created for binary segmentation tasks. The dataset is split into three parts:
146 train/val/test, in a 7:1:2 ratio. The samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values
147 between 0 and 1.

148 **DeepbacsMSBench:** The DeepbacsMSBench, based on the DeepBacs dataset^{36,37}, consists of 34 samples of fundus
149 images, each with a size of 1024×1024 pixels. It is designed for binary segmentation tasks. We use official train/validation/test
150 splits published officially by authors. The samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer
151 values between 0 and 1.

152 **DriveMSBench** The DriveMSBench dataset, based on the DRIVE dataset^{38,39}, includes 40 fundus images, each with
153 dimensions of 565×584 pixels. The images are obtained from a diabetic retinopathy screening program in the Netherlands.
154 It is designed for binary segmentation and uses official splits for training, validation, and testing. The samples are resized to
155 $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

156 **DynamicNuclearMSBench:** The DynamicNuclearMSBench, created from the DynamicNuclearNet Segmentation dataset^{40,41},
157 consists of 7084 samples of nuclear cell images, each 128×128 pixels in size. This dataset is utilized for a binary segmentation
158 task. Training, validation, and test splits that are officially published are used. The samples are resized to $1 \times 512 \times 512$ pixels,
159 and the labels are mapped to integer values between 0 and 1.

160 **FHPsAOPMSBench:** The FHPsAOPMSBench dataset is based on a prior dataset^{42,43} and comprises 4,000 ultrasound
161 images, each with a resolution of 256×256 pixels. This dataset is designed for a multi-class segmentation task, including
162 three distinct classes. The source dataset is split into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to
163 $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and (#Classes - 1).

164 **IdribMSBench:** The IdribMSBench is based on the Indian Diabetic Retinopathy Image Dataset^{44,45} and includes 80
165 high-resolution fundus images (4288×2848 pixels) for a binary segmentation task. We use official train/validation/test splits
166 published officially by authors. The authors have also categorized the labels into four categories: Microaneurysms, hemorrhages,
167 Hard Exudates, and Optic Discs. These category-based labels and annotations are provided in a npz file (see Table 3). The
168 samples are resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

169 **Isic2016MSBench:** The Isic2016MSBench is derived from the ISIC 2016 Challenge^{46,47}, which consisted of 1,279
170 dermoscopy samples of varying sizes designed for binary segmentation tasks. We use official training, validation, and test splits

171 published by authors. The samples are resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0
172 and 1.

173 **Isic2018MSBench:** The Isic2018MSBench is derived from the ISIC 2018 Challenge^{48–50}, which consisted of 3,694
174 dermoscopy samples of varying sizes designed for binary segmentation tasks. We use official training, validation, and test splits
175 published by authors. The samples are resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0
176 and 1.

177 **KvasirMSBench:** The KvasirMSBench, derived from the Kvasir-SEG dataset^{51,52}, consists of 1,000 endoscopy images
178 with resolutions ranging from 332×487 to 1920×1072 pixels. The dataset includes images of gastrointestinal polyps and
179 their segmentation masks, which have been annotated and verified by an experienced gastroenterologist. It is structured for a
180 binary classification task. The source dataset is divided into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized
181 to $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

182 **M2caiSegMSBench:** M2caiSegMSBench is based on a prior dataset^{53,54} comprising 614 pathology samples and specifically
183 designed for multi-class segmentation tasks, which include 19 distinct classes. The images within this dataset exhibit variable
184 dimensions, and we use official train/validation/test splits. The samples are resized to $3 \times 512 \times 512$ pixels, and the labels are
185 mapped to integer values between 0 and (#Classes - 1).

186 **MonusacMSBench:** MonusacMSBench is based on the MoNuSAC challenge^{55,56}. It consists of 310 samples and is
187 designed for multi-class segmentation with 6 classes. The images in this dataset are H&E stained digitized tissue images from
188 several patients acquired at multiple hospitals using a standard 40x scanner magnification. The annotations are provided by
189 expert pathologists. We use the officially published train/validation/test splits from the challenge. The samples are resized to
190 $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and (#Classes - 1).

191 **NucleiMSBench:** The NucleiMSBench is based on a prior dataset⁵⁹, which consisting of 141 pathology samples each with
192 an image size of 2000×2000 pixels. This source dataset is designed for binary segmentation tasks. The source dataset is split
193 into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to
194 integer values between 0 and 1.

195 **NusetMSBench:** The NusetMSBench, derived from the NuSet dataset^{60,61}, contains 3,408 pathology samples designed
196 for binary segmentation problems. We split the source dataset into three parts: train/val/test, in a 7:1:2 ratio. The samples are
197 resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

198 **PandentalMSBench:** The PandentalMSBench is created from the Panoramic Dental X-rays dataset^{62,63} and contains 116
199 X-ray samples of varying sizes. It is specifically intended for binary segmentation tasks. The dataset comprises anonymized
200 and deidentified panoramic dental X-rays of 116 patients taken at Noor Medical Imaging Center in Qom, Iran. The source
201 dataset is divided into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to $1 \times 512 \times 512$ pixels, and the labels
202 are mapped to integer values between 0 and 1.

203 **PolypGenMSBench:** The PolypGenMSBench is based on a prior endoscopy dataset^{64,65} consisting of 1,412 endoscopy
204 samples, each with an image size of 1920×1080 pixels. It is designed for binary segmentation tasks. It includes colonoscopy
205 video frames captured from a diverse patient population at six different centers in Egypt, France, Italy, Norway, and the United
206 Kingdom. We provide these images and annotations are captured from these centers in a npz file. The source dataset is divided
207 into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to
208 integer values between 0 and 1.

209 **Promise12MSBench:** The Promise12MSBench, derived on the PROMISE12 dataset^{66,67}, contains 1,473 MR samples,
210 each with an image size of 512×512 pixels. It is designed for binary classification. We split the source dataset into three parts:
211 train/val/test, in a 7:1:2 ratio. The samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values
212 between 0 and 1.

213 **RoboToolMSBench:** The RoboToolMSBench, based on the RoboTool dataset³¹, consisting of 500 samples, designed for
214 binary segmentation tasks. The source dataset is divided into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized
215 to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

216 **TnbcnucleiMSBench:** The TnbcnucleiMSBench is based on a prior dataset^{68,69}, consisting of 50 pathology samples,
217 each with an image size of 512×512 pixels. This dataset is based on the merging of two different datasets: the first dataset,
218 generated at the Curie Institute, consists of annotated H&E stained histology images at 40x magnification, and the second
219 dataset, provided by the Indian Institute of Technology Guwahati, also consists of annotated H&E stained histology images
220 captured at 40x magnification. It is designed for binary segmentation tasks. We split the source dataset into three parts:
221 train/val/test, in a 7:1:2 ratio. The samples are resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values
222 between 0 and 1.

223 **UltrasoundNerveMSBench:** The UltrasoundNerveMSBench, derived from prior dataset⁷⁰, contains 2,323 ultrasound
224 samples, each with an image size of 580×420 pixels and designed for binary segmentation tasks. The primary task in this
225 dataset is to segment a collection of nerves known as the Brachial Plexus (BP) in ultrasound images. Due to the lack of test

image annotations, we split the source dataset into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

USforKidneyMSBench: The USforKidneyMSBench is derived from the CT2USforKidneySeg dataset^{71,72}, comprised of 4,586 ultrasound samples, each with an image size of 256×256 pixels, and designed for binary segmentation tasks. The source dataset is split into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

UWSkinCancerMSBench: The UWSkinCancerMSBench is based on the Skin Cancer Detection dataset⁷³, consisting of 206 dermoscopy samples, designed for binary classification tasks. The dataset includes images extracted from the public databases DermIS and DermQuest, along with manual segmentations of the lesions. We split the source dataset into three parts: train/val/test, in a 7:1:2 ratio. The authors have also categorized the labels into two categories: Melanoma and Not-Melanoma. These category-based labels and annotations are provided in a .npz file (see Table 3). The samples are resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

WbcMSBench: The WbcMSBench, based on prior datasets^{74,75}, is a microscopy imaging dataset consisting of 80 samples, with image sizes of 120×120 and 300×300 pixels. It is designed for multi-class segmentation tasks including 3 classes. The dataset is based on two sources: Dataset 1, obtained from Jiangxi Tecom Science Corporation, China, contains 300 images of white blood cells with a resolution of 120×120 pixels. Dataset 2 consists of 100 color images with a resolution of 300×300 pixels, collected from the CellaVision blog. The authors have grouped the samples into four categories: Lymphocyte, Monocyte, Neutrophil, and Eosinophil, and we provide these category-based images and corresponding labels in npz file format (see Table 3). The source dataset is divided into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to $3 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and (#Classes - 1).

YeazMSBench: The YeazMSBench, derived from the YeaZ dataset^{76,77}, consists of 707 microscopy images with varying sizes and is designed for binary segmentation tasks. We split the source dataset into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to $1 \times 512 \times 512$ pixels, and the labels are mapped to integer values between 0 and 1.

Data Records

We have publicly shared each dataset with varying sizes (128, 256, and 512 sized) in MedSegBench at Zenodo ([Link](#)). The MedSegBench consists of 35 pre-processed 2D medical image segmentation datasets (some of them extracted 3D slices) from various data modalities and tasks (binary/multi-class). The data storage format published by MedMNISTv2⁸ is followed. We save each dataset in Numpy npz format, named as {dataset}_{size}.npz. Each npz file contains following keys: [{"train,val,test}_images", {"train,val,test}_label"]. Also, some authors have published class- or category-based images and labels. We have also added this information with the following keys into the npz file and explain them in source code files: [{"train,val,test}_images_{classno}", {"train,val,test}_label_{classno}"]. All images and labels are stored in uint8 data type. **{train,val,test}_images:** Numpy array contains train, validation and test images with $N \times W \times H \times C$ shape for RGB datasets, and $N \times W \times H$ for gray-scale datasets. Here, N refers to the number of samples, W is the width, H is the height, and C denotes the number of channels. **{train,val,test}_label:** It contains train, validation and test labels with $N \times W \times H$ shape. **{train,val,test}_images_{classno}** and **{train,val,test}_label_{classno}**: These contain class or category-based train, validation, and test images and labels with shapes $N \times W \times H \times C$ (for RGB images, and $N \times W \times H$ for gray-scale images), respectively.

Technical Validation

Baseline methods

In this study, we chose the U-Net architecture as the baseline structure for image segmentation tasks. We have selected six encoder/decoder networks to enhance performance and adaptability. These include ResNet18, ResNet50, and DenseNet121, commonly used as benchmarks in segmentation research. Additionally, we have selected EfficientNet and MobileNetv2 because they are lightweight models that offer a more computationally efficient alternative to ResNets and DenseNet. Furthermore, we have added a transformer-based approach using the Mix Vision Transformer, acknowledging the growing interest in transformer models for vision tasks.

The U-Net structure and diverse encoder/decoder networks are implemented using the qubvel-segmentation framework¹⁰. We have not used pre-trained ImageNet weights; we train each model from scratch on our datasets. We have trained each model with three randomly selected seed values to ensure the robustness of our results. All images are resized to 512×512 pixels, a standardized dimension for the training, validation, and testing phases. Training is conducted over 200 epochs using the Adam Optimizer with a learning rate $1e-3$. For binary segmentation tasks, we used dice loss, while categorical cross-entropy loss is used for multi-class tasks. A batch size of 128 is selected throughout the training process. We have not applied weight decay methods or any data augmentation techniques, focusing on the raw performance of the models. The model weights

278 corresponding to the best validation IOU are recorded for each network configuration. Further details regarding the model
279 implementation, training, and evaluation steps are available in our code repository.

280 Performance Measures

281 We have evaluated each model on 35 different datasets using four performance measures: Precision (PREC), Recall (REC),
282 F1-score (F1), and Intersection over Union (IOU). Precision measures the accuracy of positive predictions, highlighting its
283 ability to avoid false positives, while Recall evaluates the model's capacity to identify all relevant positive instances, minimizing
284 false negatives. The F1-Score, as the harmonic mean of Precision and Recall, provides a balanced view, which is especially
285 useful when there is an unbalanced class distribution. IoU, primarily used in image segmentation and object detection, evaluates
286 the overlap between predicted and actual regions, ensuring accurate localization and identification of objects. We have
287 individually reported PREC, REC, F1, and IOU scores for each dataset and averaged the results.

288 Results

289 The average PREC and REC results obtained from three different run are showed in Table 4 and average F1 and IOU scores are
290 reported in Table 5 for each individual datasets. Also, the average results for each baseline methods are shown in Table 5,

291 Table 4 presents a comprehensive overview of the average precision and recall results for six different encoder networks
292 across various datasets. These networks include ResNet-18 (RN-18), ResNet-50 (RN-50), Efficient-Net (EN), Mobile-Net-v2
293 (MN-v2), DenseNet-121 (DN-121), and Mix Vision Transformer (MVT). The results are divided into two main categories:
294 precision and recall. In terms of precision, DenseNet-121 consistently demonstrated strong performance across numerous
295 datasets. For example, it achieved the highest precision scores in datasets such as BusiMSB (0.794), ChuahMSB(0.870) and
296 Dca1MSB (0.801). Similarly, Efficient-Net also demonstrated strong precision, particularly in datasets like Isic2016MSB and
297 Isic2018MSB, where it scored 0.912 and 0.857, respectively. Although the Mix Vision Transformer is not evaluated on all
298 datasets because it only accepts at least three channel images as input, it performed competitively where applicable, achieving
299 high precision in datasets like Bkai-Igh-MSB (0.983). Regarding Recall, DenseNet-121 has emerged as a top performer,
300 achieving the highest recall in datasets such as Bbbbc010MSB (0.920) and WbcMSB (0.970). Efficient-Net also performed well
301 in recall metric, particularly in datasets like DynamicNuclearMSB (0.966) and USforKidneyMSB (0.982). The results indicate
302 that DenseNet-121 and Efficient-Net are particularly robust across precision and recall metrics, suggesting their effectiveness in
303 various applications. Overall, the analysis highlights DenseNet-121's consistently high performance across multiple datasets,
304 making it a reliable choice for tasks requiring high precision and recall. Efficient-Net also stands out, especially in recall,
305 indicating its potential for applications where recall is critical.

306 Table 5 provides a comprehensive evaluation of six encoder networks across various datasets, using F1-score and Intersection
307 over Union (IOU) as performance metrics. DenseNet-121 consistently performs well, frequently achieving the top F1 and
308 IOU metrics scores across numerous datasets. For example, in the Bbbbc010MSB and CellNucleiMSB datasets, DenseNet-121
309 records the highest F1-scores of 0.920 and 0.907, respectively, and similarly high IOU scores, indicating its robustness in
310 handling diverse data types. Efficient-Net also shows significant performance, particularly in datasets like Isic2016MSB and
311 USforKidneyMSB, where it achieves the highest F1-scores of 0.903 and 0.981, respectively. This indicates that Efficient-Net is
312 particularly effective in scenarios requiring high precision and recall, as showed in its F1-scores. ResNet-50 performs best
313 with an F1-score of 0.931 and an IOU of 0.870 for the DeepbacsMSB. Additionally, it has also achieved the highest F1-score
314 of 0.786 and an IOU of 0.648 in the DriveMSB dataset. For the FHPsAOPMSB dataset, ResNet-18 has achieved the highest
315 F1-score of 0.961 and an IOU of 0.929. While Mix Vision Transformer does not frequently perform as well as DenseNet-121,
316 it shows competitive performance in specific datasets such as UWSkinCancerMSB, achieving the second-highest F1 Score of
317 0.881. This indicates its potential in specialized applications, particularly in medical imaging contexts. Overall, DenseNet-121
318 is the most robust and effective network, frequently outperforming other networks in achieving high F1-scores and IOU values.

319 Table 6 shows the average performance metrics for six different encoder networks. Efficient-Net (EN) and DenseNet-121
320 (DN-121) demonstrate the highest F1 scores, both achieving a value of 0.772. This indicates that these models have a balanced
321 performance in terms of precision and recall. DenseNet-121 also achieves the highest precision at 0.848, suggesting it is
322 effective at minimizing false positives. On the other hand, Efficient-Net leads in recall with a score of 0.788, indicating its
323 strength in capturing true positives. Additionally, DenseNet-121 achieves the highest IOU with 0.702, closely followed by
324 Efficient-Net with 0.700 This suggests that these two models provide the most accurate predictions. Overall, DenseNet-121 and
325 Efficient-Net achieve similar high-performance metrics, with both models performing well in F1 score, precision, recall, and
326 IOU. However, DenseNet-121's complex architecture causes higher computational demands, whereas Efficient-Net provides a
327 more efficient design, making it suitable for resource-constrained applications.

Table 4. The average precision and recall results for six different encoder networks. RN-18: ResNet-18; RN-50: ResNet-50; EN: Efficient-Net; MN-v2: Mobile-Net-v2; DN-121: DenseNet-121; MVT: Mix Vision Transformer

Dataset	Precision (PREC)						Recall (REC)					
	RN-18	RN-50	EN	MN-v2	DN-121	MVT	RN-18	RN-50	EN	MN-v2	DN-121	MVT
AbdomenUSMSB	0.976	0.973	0.950	0.964	0.955	-	0.652	0.654	0.670	0.655	0.671	-
Bbbc010MSB	0.919	0.926	0.918	0.918	0.922	-	0.912	0.909	0.904	0.900	0.920	-
Bkai-Igh-MSB	0.983	0.961	0.939	0.944	0.952	0.983	0.563	0.625	0.705	0.737	0.642	0.563
BriFiSegMSB	0.812	0.816	0.812	0.803	0.817	-	0.873	0.886	0.882	0.861	0.898	-
BusiMSB	0.729	0.753	0.765	0.766	0.794	-	0.727	0.665	0.728	0.672	0.714	-
CellNucleiMSB	0.924	0.920	0.913	0.901	0.927	0.928	0.882	0.886	0.894	0.872	0.898	0.883
ChaseDB1MSB	0.788	0.789	0.780	0.794	0.793	0.774	0.733	0.738	0.725	0.703	0.739	0.705
ChuacMSB	0.713	0.710	0.643	0.644	0.870	-	0.470	0.451	0.526	0.458	0.444	-
Covid19RadioMSB	0.991	0.991	0.991	0.991	0.992	-	0.990	0.990	0.991	0.991	0.991	-
CovidQUExMSB	0.741	0.738	0.753	0.739	0.760	-	0.824	0.810	0.815	0.827	0.826	-
CystoFluidMSB	0.889	0.870	0.874	0.879	0.888	0.874	0.848	0.872	0.856	0.844	0.851	0.865
Dca1MSB	0.776	0.788	0.775	0.781	0.801	-	0.757	0.757	0.740	0.732	0.740	-
DeepbacsMSB	0.957	0.956	0.955	0.958	0.959	-	0.905	0.907	0.897	0.886	0.900	-
DriveMSB	0.817	0.789	0.799	0.811	0.827	0.784	0.756	0.790	0.748	0.750	0.751	0.784
DynamicNuclearMSB	0.924	0.929	0.937	0.926	0.928	-	0.965	0.965	0.966	0.963	0.965	-
FHPsAOPMSB	0.962	0.964	0.964	0.965	0.961	-	0.960	0.951	0.956	0.955	0.959	-
IdribMSB	0.150	0.153	0.139	0.150	0.172	0.110	0.089	0.072	0.065	0.078	0.068	0.041
Isic2016MSB	0.890	0.897	0.912	0.912	0.913	0.897	0.907	0.910	0.919	0.901	0.905	0.917
Isic2018MSB	0.838	0.839	0.857	0.864	0.878	0.854	0.911	0.907	0.923	0.908	0.896	0.907
KvasirMSB	0.816	0.770	0.839	0.842	0.874	0.644	0.768	0.755	0.860	0.780	0.804	0.697
M2caiSegMSB	0.737	0.756	0.801	0.762	0.759	0.794	0.224	0.225	0.228	0.225	0.230	0.227
MonusacMSB	0.945	0.951	0.951	0.951	0.951	0.951	0.951	0.589	0.589	0.589	0.589	0.589
MosMedPlusMSB	0.816	0.817	0.807	0.821	0.826	0.808	0.786	0.802	0.796	0.793	0.798	0.767
NucleiMSB	0.250	0.233	0.223	0.199	0.225	0.196	0.394	0.395	0.449	0.281	0.479	0.481
NusetMSB	0.949	0.950	0.953	0.950	0.953	-	0.951	0.951	0.951	0.952	0.952	-
PandentalMSB	0.956	0.955	0.952	0.945	0.965	-	0.967	0.968	0.963	0.958	0.965	-
PolypGenMSB	0.763	0.739	0.783	0.824	0.794	0.557	0.584	0.538	0.684	0.582	0.632	0.570
Promise12MSB	0.911	0.900	0.900	0.903	0.909	-	0.903	0.896	0.902	0		
RoboToolMSB	0.878	0.874	0.893	0.885	0.905	0.885	0.854	0.864	0.867	0.835	0.868	0.893
TnbnucleiMSB	0.813	0.834	0.748	0.772	0.819	0.746	0.758	0.760	0.762	0.770	0.770	0.797
UltrasoundNerveMSB	0.799	0.801	0.779	0.786	0.798	-	0.796	0.782	0.814	0.791	0.802	-
USforKidneyMSB	0.979	0.979	0.981	0.980	0.980	-	0.980	0.978	0.982	0.980	0.980	-
UWSkinCancerMSB	0.920	0.925	0.928	0.939	0.926	0.930	0.857	0.829	0.882	0.857	0.839	0.872
WbcMSB	0.961	0.962	0.965	0.959	0.963	0.966	0.966	0.966	0.968	0.963	0.970	0.969
YeazMSB	0.935	0.931	0.936	0.931	0.934	-	0.974	0.979	0.971	0.977	0.978	-

Table 5. The average F1-score and IOU results for six different encoder networks. RN-18: ResNet-18; RN-50: ResNet-50; EN: Efficient-Net; MN-v2: Mobile-Net-v2; DN-121: DenseNet-121; MVT: Mix Vision Transformer

Dataset	F1-Score (F1)						Intersection over Union (IOU)					
	RN-18	RN-50	EN	MN-v2	DN-121	MVT	RN-18	RN-50	EN	MN-v2	DN-121	MVT
AbdomenUSMSB	0.642	0.640	0.640	0.635	0.643	-	0.632	0.630	0.628	0.624	0.632	-
Bbbc010MSB	0.915	0.917	0.910	0.908	0.920	-	0.844	0.848	0.837	0.833	0.854	-
Bkai-Igh-MSB	0.554	0.617	0.692	0.733	0.630	0.554	0.546	0.604	0.676	0.713	0.615	0.546
BriFiSegMSB	0.826	0.834	0.831	0.816	0.840	-	0.717	0.728	0.724	0.702	0.738	-
BusiMSB	0.674	0.632	0.711	0.655	0.695	-	0.578	0.547	0.624	0.565	0.615	-
CellNucleiMSB	0.889	0.892	0.894	0.880	0.907	0.891	0.822	0.827	0.830	0.815	0.838	0.826
ChaseDB1MSB	0.758	0.761	0.750	0.744	0.764	0.735	0.611	0.615	0.601	0.594	0.618	0.582
ChuacMSB	0.487	0.451	0.499	0.462	0.522	-	0.357	0.334	0.369	0.340	0.400	-
Covid19RadioMSB	0.991	0.990	0.991	0.991	0.992	-	0.982	0.981	0.983	0.982	0.983	-
CovidQUExMSB	0.740	0.734	0.744	0.742	0.756	-	0.627	0.620	0.633	0.631	0.647	-
CystoFluidMSB	0.852	0.857	0.849	0.842	0.853	0.855	0.759	0.765	0.754	0.747	0.761	0.763
Dca1MSB	0.762	0.767	0.753	0.751	0.765	-	0.618	0.625	0.606	0.604	0.623	-
DeepbacsMSB	0.930	0.931	0.925	0.921	0.929	-	0.869	0.870	0.860	0.853	0.867	-
DriveMSB	0.782	0.786	0.770	0.775	0.782	0.781	0.643	0.648	0.626	0.634	0.643	0.641
DynamicNuclearMSB	0.941	0.942	0.948	0.940	0.942	-	0.895	0.897	0.906	0.893	0.897	-
FHPsAOPMSB	0.961	0.957	0.959	0.959	0.960	-	0.929	0.923	0.927	0.927	0.928	-
IdribMSB	0.100	0.090	0.078	0.092	0.089	0.053	0.061	0.054	0.046	0.056	0.054	0.030
Isic2016MSB	0.878	0.887	0.903	0.891	0.893	0.891	0.803	0.814	0.836	0.820	0.825	0.822
Isic2018MSB	0.849	0.849	0.868	0.865	0.861	0.853	0.761	0.762	0.790	0.783	0.785	0.773
KvasirMSB	0.739	0.698	0.812	0.754	0.794	0.569	0.645	0.596	0.733	0.668	0.718	0.457
M2caiSegMSB	0.214	0.215	0.218	0.216	0.223	0.217	0.190	0.191	0.196	0.192	0.200	0.194
MonusacMSB	0.557	0.559	0.559	0.559	0.559	0.538	0.540	0.540	0.540	0.540	0.540	0.540
MosMedPlusMSB	0.780	0.790	0.781	0.785	0.791	0.761	0.674	0.682	0.674	0.679	0.686	0.650
NucleiMSB	0.282	0.274	0.278	0.205	0.275	0.253	0.169	0.164	0.167	0.119	0.166	0.150
NusetMSB	0.949	0.949	0.951	0.950	0.951	-	0.906	0.906	0.909	0.907	0.910	-
PandentalMSB	0.961	0.961	0.957	0.950	0.965	-	0.926	0.926	0.919	0.907	0.932	-
PolypGenMSB	0.573	0.541	0.666	0.588	0.621	0.477	0.495	0.457	0.587	0.512	0.545	0.382
Promise12MSB	0.895	0.888	0.892	0.896	0.900	-	0.828	0.817	0.821	0.827	0.832	-
RoboToolMSB	0.856	0.859	0.874	0.847	0.879	0.882	0.765	0.769	0.788	0.753	0.798	0.798
TnbucleiMSB	0.779	0.785	0.738	0.762	0.788	0.759	0.641	0.652	0.596	0.621	0.654	0.618
UltrasoundNerveMSB	0.782	0.776	0.787	0.772	0.786	-	0.671	0.664	0.675	0.660	0.676	-
USforKidneyMSB	0.979	0.978	0.981	0.980	0.980	-	0.960	0.958	0.963	0.961	0.960	-
UWSkinCancerMSB	0.864	0.846	0.890	0.879	0.856	0.881	0.795	0.766	0.818	0.803	0.779	0.813
WbcMSB	0.962	0.963	0.966	0.959	0.966	0.967	0.930	0.931	0.937	0.926	0.936	0.938
YeazMSB	0.953	0.953	0.952	0.952	0.954	-	0.912	0.912	0.909	0.910	0.914	-

Table 6. The average results for six different encoder networks. RN-18: ResNet-18; RN-50: ResNet-50; EN: Efficient-Net; MN-v2: Mobile-Net-v2; DN-121: DenseNet-121; MVT: Mix Vision Transformer.

Methods	F1	PREC	REC	IOU
RN-18	0.762	0.834	0.774	0.689
RN-50	0.759	0.833	0.772	0.686
EN	0.772	0.832	0.788	0.700
MN-v2	0.762	0.834	0.769	0.689
DN-121	0.772	0.848	0.781	0.702
MVT	0.663	0.760	0.696	0.585

328 Usage Notes

329 The MedSegBench datasets are freely available at [Zenodo](https://zenodo.org/). We kindly request that users of the MedSegBench dataset cite
330 this paper, along with the relevant source dataset files, in their publications. This dataset is created in order to fairly compare
331 different models over various segmentation models from different data modalities and to create universal models. It is not
332 suitable for clinical or medical use.

333 Code availability

334 The Python data API, source code files and evaluation scripts for binary and multi-class segmentation tasks can be found at
335 <https://github.com/zekikus/MedSegBench>.

336 References

- 337 1. Han, K. *et al.* Deep semi-supervised learning for medical image segmentation: A review. *Expert. Syst. with Appl.* **245**,
338 123052, [10.1016/j.eswa.2023.123052](https://doi.org/10.1016/j.eswa.2023.123052) (2024).
- 339 2. Ma, J. *et al.* Segment anything in medical images. *Nat. Commun.* **15**, [10.1038/s41467-024-44824-z](https://doi.org/10.1038/s41467-024-44824-z) (2024).
- 340 3. Carriero, A., Groenhoff, L., Vologina, E., Basile, P. & Albera, M. Deep learning in breast cancer imaging: State of the art
341 and recent advancements in early 2024. *Diagnostics* **14**, 848, [10.3390/diagnostics14080848](https://doi.org/10.3390/diagnostics14080848) (2024).
- 342 4. Drelie Gelasca, E., Obara, B., Fedorov, D., Kvilekval, K. & Manjunath, B. A biosegmentation benchmark for evaluation of
343 bioimage analysis methods. *BMC Bioinforma.* **10**, [10.1186/1471-2105-10-368](https://doi.org/10.1186/1471-2105-10-368) (2009).
- 344 5. Rebuffi, S.-A., Bilen, H. & Vedaldi, A. Learning multiple visual domains with residual adapters. In Guyon, I. *et al.* (eds.)
345 *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017).
- 346 6. Simpson, A. L. *et al.* A large annotated medical image dataset for the development and evaluation of segmentation
347 algorithms. *CoRR abs/1902.09063* (2019). [1902.09063](https://arxiv.org/abs/1902.09063).
- 348 7. Yang, J., Shi, R. & Ni, B. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis.
349 In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, [10.1109/isbi48211.2021.9434062](https://doi.org/10.1109/isbi48211.2021.9434062) (IEEE,
350 2021).
- 351 8. Yang, J. *et al.* Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Sci. Data*
352 **10**, [10.1038/s41597-022-01721-8](https://doi.org/10.1038/s41597-022-01721-8) (2023).
- 353 9. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, 234–241
354 (Springer International Publishing, 2015).
- 355 10. Iakubovskii, P. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch (2019).
- 356 11. Vitale, S., Orlando, J. I., Iarussi, E. & Larrabide, I. Improving realism in patient-specific abdominal ultrasound simulation
357 using cyclegans. *Int. J. Comput. Assist. Radiol. Surg.* **15**, 183–192, [10.1007/s11548-019-02046-5](https://doi.org/10.1007/s11548-019-02046-5) (2019).
- 358 12. Orlando, J. I. Us simulation & segmentation (2020).
- 359 13. Ljosa, V., Sokolnicki, K. L. & Carpenter, A. E. Annotated high-throughput microscopy image sets for validation. *Nat.*
360 *Methods* **9**, 637–637, [10.1038/nmeth.2083](https://doi.org/10.1038/nmeth.2083) (2012).
- 361 14. Broad Bioimage Benchmark Collection — bbbc.broadinstitute.org. <https://bbbc.broadinstitute.org/BBBC010>. [Accessed
362 06-08-2024].
- 363 15. Ngoc Lan, P. *et al.* *NeoUNet: Towards Accurate Colon Polyp Segmentation and Neoplasm Detection*, 15–28 (Springer
364 International Publishing, 2021).
- 365 16. An, N. S. *et al.* Blazeneo: Blazing fast polyp segmentation and neoplasm detection. *IEEE Access* **10**, 43669–43684,
366 [10.1109/access.2022.3168693](https://doi.org/10.1109/access.2022.3168693) (2022).
- 367 17. Duc, N. T., Oanh, N. T., Thuy, N. T., Triet, T. M. & Dinh, V. S. Colonformer: An efficient transformer based method for
368 colon polyp segmentation. *IEEE Access* **10**, 80575–80586, [10.1109/access.2022.3195241](https://doi.org/10.1109/access.2022.3195241) (2022).
- 369 18. Mathieu, G., M., L. A. & Bachir, E. D. Brifiseg: a deep learning-based method for semantic and instance segmentation of
370 nuclei in brightfield images, [10.48550/ARXIV.2211.03072](https://doi.org/10.48550/ARXIV.2211.03072) (2022).
- 371 19. Gendarme, M. & Debs, B. E. Brifiseg datasets, [10.5281/ZENODO.7195636](https://zenodo.org/record/7195636) (2022).
- 372 20. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data Brief* **28**, 104863,
373 [10.1016/j.dib.2019.104863](https://doi.org/10.1016/j.dib.2019.104863) (2020).

- 374 **21.** Breast Ultrasound Images Dataset — kaggle.com. [https://www.kaggle.com/datasets/aryashah2k/](https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset)
375 [breast-ultrasound-images-dataset](https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset). [Accessed 06-08-2024].
- 376 **22.** Caicedo, J. C. *et al.* Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nat. Methods* **16**,
377 1247–1253, [10.1038/s41592-019-0612-7](https://doi.org/10.1038/s41592-019-0612-7) (2019).
- 378 **23.** 2018 Data Science Bowl — kaggle.com. <https://www.kaggle.com/competitions/data-science-bowl-2018/data>. [Accessed
379 06-08-2024].
- 380 **24.** Carballal, A. *et al.* Automatic multiscale vascular image segmentation algorithm for coronary angiography. *Biomed. Signal*
381 *Process. Control.* **46**, 1–9, [10.1016/j.bspc.2018.06.007](https://doi.org/10.1016/j.bspc.2018.06.007) (2018).
- 382 **25.** Angiographics — figshare.com. <https://figshare.com/s/4d24cf3d14bc901a94bf>. [Accessed 06-08-2024].
- 383 **26.** Chowdhury, M. E. H. *et al.* Can ai help in screening viral and covid-19 pneumonia? *IEEE Access* **8**, 132665–132676,
384 [10.1109/access.2020.3010287](https://doi.org/10.1109/access.2020.3010287) (2020).
- 385 **27.** Rahman, T. *et al.* Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images.
386 *Comput. Biol. Medicine* **132**, 104319, [10.1016/j.combiomed.2021.104319](https://doi.org/10.1016/j.combiomed.2021.104319) (2021).
- 387 **28.** COVID-19 Radiography Database — kaggle.com. [https://www.kaggle.com/datasets/tawsifurrahman/](https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database)
388 [covid19-radiography-database](https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database). [Accessed 06-08-2024].
- 389 **29.** Tahir, A. M. *et al.* Covid-19 infection localization and severity grading from chest x-ray images. *Comput. Biol. Medicine*
390 **139**, 105002, [10.1016/j.combiomed.2021.105002](https://doi.org/10.1016/j.combiomed.2021.105002) (2021).
- 391 **30.** Anas M. Tahir *et al.* Covid-qu-ex dataset, [10.34740/KAGGLE/DSV/3122958](https://doi.org/10.34740/KAGGLE/DSV/3122958) (2022).
- 392 **31.** Garcia-Peraza-Herrera, L. C. *et al.* Image compositing for segmentation of surgical tools without manual annotations.
393 *IEEE Transactions on Med. Imaging* **40**, 1450–1460, [10.1109/tmi.2021.3057884](https://doi.org/10.1109/tmi.2021.3057884) (2021).
- 394 **32.** Zeeshan Ahmed, Munawar Ahmed, Attiya Baqai & Fahim Aziz Umrani. Intraretinal cystoid fluid, [10.34740/KAGGLE/](https://doi.org/10.34740/KAGGLE/DS/2277068)
395 [DS/2277068](https://doi.org/10.34740/KAGGLE/DS/2277068) (2022).
- 396 **33.** Ahmed, Z. *et al.* Deep learning based automated detection of intraretinal cystoid fluid. *Int. J. Imaging Syst. Technol.* **32**,
397 902–917, [10.1002/ima.22662](https://doi.org/10.1002/ima.22662) (2021).
- 398 **34.** Cervantes-Sanchez, F., Cruz-Aceves, I., Hernandez-Aguirre, A., Hernandez-Gonzalez, M. A. & Solorio-Meza, S. E.
399 Automatic segmentation of coronary arteries in x-ray angiograms using multiscale analysis and artificial neural networks.
400 *Appl. Sci.* **9**, 5507, [10.3390/app9245507](https://doi.org/10.3390/app9245507) (2019).
- 401 **35.** Ivan Cruz Aceves CIMAT — personal.cimat.mx. http://personal.cimat.mx:8181/~ivan.cruz/DB_Angiograms.html. [Ac-
402 cessed 06-08-2024].
- 403 **36.** Spahn, C. *et al.* Deepbacs for multi-task bacterial image analysis using open-source deep learning approaches. *Commun.*
404 *Biol.* **5**, [10.1038/s42003-022-03634-z](https://doi.org/10.1038/s42003-022-03634-z) (2022).
- 405 **37.** Spahn, C. & Heilemann, M. Deepbacs – escherichia coli bright field segmentation dataset, [10.5281/ZENODO.5550934](https://doi.org/10.5281/ZENODO.5550934)
406 (2021).
- 407 **38.** Staal, J., Abramoff, M., Niemeijer, M., Viergever, M. & van Ginneken, B. Ridge-based vessel segmentation in color images
408 of the retina. *IEEE Transactions on Med. Imaging* **23**, 501–509, [10.1109/tmi.2004.825627](https://doi.org/10.1109/tmi.2004.825627) (2004).
- 409 **39.** DRIVE - Grand Challenge — drive.grand-challenge.org. <https://drive.grand-challenge.org/>. [Accessed 06-08-2024].
- 410 **40.** Van Valen, D. A. *et al.* Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments.
411 *PLOS Comput. Biol.* **12**, e1005177, [10.1371/journal.pcbi.1005177](https://doi.org/10.1371/journal.pcbi.1005177) (2016).
- 412 **41.** DeepCell Datasets — datasets.deepcell.org. <https://datasets.deepcell.org/data>. [Accessed 06-08-2024].
- 413 **42.** Lu, Y. *et al.* The jnu-ifm dataset for segmenting pubic symphysis-fetal head. *Data Brief* **41**, 107904, [10.1016/j.dib.2022.](https://doi.org/10.1016/j.dib.2022.107904)
414 [107904](https://doi.org/10.1016/j.dib.2022.107904) (2022).
- 415 **43.** Jieyun, B. & ZhanHong, O. Pubic symphysis-fetal head segmentation and angle of progression, [10.5281/ZENODO.7851338](https://doi.org/10.5281/ZENODO.7851338)
416 (2024).
- 417 **44.** Porwal, P. *et al.* Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research.
418 *Data* **3**, 25, [10.3390/data3030025](https://doi.org/10.3390/data3030025) (2018).
- 419 **45.** Prasanna Porwal, S. P. Indian diabetic retinopathy image dataset (idrid), [10.21227/H25W98](https://doi.org/10.21227/H25W98) (2018).

- 420 **46.** Codella, N. C. F. *et al.* Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium
421 on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International*
422 *Symposium on Biomedical Imaging (ISBI 2018)*, [10.1109/isbi.2018.8363547](https://doi.org/10.1109/isbi.2018.8363547) (IEEE, 2018).
- 423 **47.** ISIC Challenge — challenge.isic-archive.com. <https://challenge.isic-archive.com/data/#2016>. [Accessed 07-08-2024].
- 424 **48.** Tschandl, P., Rosendahl, C. & Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images
425 of common pigmented skin lesions. *Sci. Data* **5**, [10.1038/sdata.2018.161](https://doi.org/10.1038/sdata.2018.161) (2018).
- 426 **49.** Codella, N. *et al.* Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin
427 imaging collaboration (isic), [10.48550/ARXIV.1902.03368](https://doi.org/10.48550/ARXIV.1902.03368) (2019).
- 428 **50.** ISIC Challenge — challenge.isic-archive.com. <https://challenge.isic-archive.com/data/#2018>. [Accessed 07-08-2024].
- 429 **51.** Jha, D. *et al.* *Kvasir-SEG: A Segmented Polyp Dataset*, 451–462 (Springer International Publishing, 2019).
- 430 **52.** Simula Datasets - Kvasir SEG — datasets.simula.no. <https://datasets.simula.no/kvasir-seg/>. [Accessed 06-08-2024].
- 431 **53.** Maqbool, S., Riaz, A., Sajid, H. & Hasan, O. m2caiseg: Semantic segmentation of laparoscopic images using convolutional
432 neural networks, [10.48550/ARXIV.2008.10134](https://doi.org/10.48550/ARXIV.2008.10134) (2020).
- 433 **54.** m2caiSeg — kaggle.com. <https://www.kaggle.com/datasets/salmanmaq/m2caiseg>. [Accessed 07-08-2024].
- 434 **55.** Verma, R. *et al.* Monusac2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Transactions on*
435 *Med. Imaging* **40**, 3413–3423, [10.1109/tmi.2021.3085712](https://doi.org/10.1109/tmi.2021.3085712) (2021).
- 436 **56.** MoNuSAC 2020 - Grand Challenge — monusac-2020.grand-challenge.org. [https://monusac-2020.grand-challenge.org/](https://monusac-2020.grand-challenge.org/Data/)
437 [Data/](https://monusac-2020.grand-challenge.org/Data/). [Accessed 07-08-2024].
- 438 **57.** Morozov, S. P. *et al.* Mosmeddata: Chest ct scans with covid-19 related findings dataset, [10.48550/ARXIV.2005.06465](https://doi.org/10.48550/ARXIV.2005.06465)
439 (2020).
- 440 **58.** COVID-19 CT scan lesion segmentation dataset — kaggle.com. [https://www.kaggle.com/datasets/maedemaftouni/](https://www.kaggle.com/datasets/maedemaftouni/covid19-ct-scan-lesion-segmentation-dataset)
441 [covid19-ct-scan-lesion-segmentation-dataset](https://www.kaggle.com/datasets/maedemaftouni/covid19-ct-scan-lesion-segmentation-dataset). [Accessed 06-08-2024].
- 442 **59.** Janowczyk, A. & Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with
443 selected use cases. *J. Pathol. Informatics* **7**, 29, [10.4103/2153-3539.186902](https://doi.org/10.4103/2153-3539.186902) (2016).
- 444 **60.** Yang, L. *et al.* Nuset: A deep learning tool for reliably separating and analyzing crowded cells. *PLOS Comput. Biol.* **16**,
445 e1008193, [10.1371/journal.pcbi.1008193](https://doi.org/10.1371/journal.pcbi.1008193) (2020).
- 446 **61.** Linfeng Yang. Nuset training dataset/model weights from (nuset: A deep learning tool for reliably separating and analyzing
447 crowded cells), [10.5281/ZENODO.3996369](https://doi.org/10.5281/ZENODO.3996369) (2020).
- 448 **62.** Abdi, A. H., Kasaei, S. & Mehdizadeh, M. Automatic segmentation of mandible in panoramic x-ray. *J. Med. Imaging* **2**,
449 044003, [10.1117/1.jmi.2.4.044003](https://doi.org/10.1117/1.jmi.2.4.044003) (2015).
- 450 **63.** Abdi, A. Panoramic dental x-rays with segmented mandibles, [10.17632/HXT48YK462.1](https://doi.org/10.17632/HXT48YK462.1) (2017).
- 451 **64.** Ali, S. *et al.* Assessing generalisability of deep learning-based polyp detection and segmentation methods through a
452 computer vision challenge, [10.48550/ARXIV.2202.12031](https://doi.org/10.48550/ARXIV.2202.12031) (2022).
- 453 **65.** Ali, S. *et al.* Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy.
454 *Med. Image Analysis* **70**, 102002, [10.1016/j.media.2021.102002](https://doi.org/10.1016/j.media.2021.102002) (2021).
- 455 **66.** Litjens, G. *et al.* Evaluation of prostate segmentation algorithms for mri: The promise12 challenge. *Med. Image Analysis*
456 **18**, 359–373, [10.1016/j.media.2013.12.002](https://doi.org/10.1016/j.media.2013.12.002) (2014).
- 457 **67.** Litjens, G. *et al.* Promise12: Data from the miccai grand challenge: Prostate mr image segmentation 2012, [10.5281/](https://doi.org/10.5281/ZENODO.8014040)
458 [ZENODO.8014040](https://doi.org/10.5281/ZENODO.8014040) (2023).
- 459 **68.** Jack, N. P., Thomas, W., Laé Marick & Reyat Fabien. Segmentation of nuclei in histopathology images by deep regression
460 of the distance map, [10.5281/ZENODO.1175282](https://doi.org/10.5281/ZENODO.1175282) (2018).
- 461 **69.** Naylor, P., Laé, M., Reyat, F. & Walter, T. Segmentation of nuclei in histopathology images by deep regression of the
462 distance map. *IEEE Transactions on Med. Imaging* **38**, 448–459, [10.1109/tmi.2018.2865709](https://doi.org/10.1109/tmi.2018.2865709) (2019).
- 463 **70.** Ultrasound Nerve Segmentation — kaggle.com. <https://www.kaggle.com/competitions/ultrasound-nerve-segmentation>.
464 [Accessed 07-08-2024].
- 465 **71.** Song, Y. *et al.* Ct2us: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data.
466 *Ultrasonics* **122**, 106706, [10.1016/j.ultras.2022.106706](https://doi.org/10.1016/j.ultras.2022.106706) (2022).

- 467 **72.** CT2USforKidneySeg — kaggle.com. <https://www.kaggle.com/datasets/siatsyx/ct2usforkidneyseg/data>. [Accessed 07-08-
468 2024].
- 469 **73.** Skin Cancer Detection | Vision and Image Processing Lab — uwaterloo.ca. [https://uwaterloo.ca/
470 vision-image-processing-lab/research-demos/skin-cancer-detection](https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection). [Accessed 07-08-2024].
- 471 **74.** Zheng, X., Wang, Y., Wang, G. & Liu, J. Fast and robust segmentation of white blood cell images by self-supervised
472 learning. *Micron* **107**, 55–71, [10.1016/j.micron.2018.01.010](https://doi.org/10.1016/j.micron.2018.01.010) (2018).
- 473 **75.** Acevedo, A., Alférez, S., Merino, A., Puigví, L. & Rodellar, J. Recognition of peripheral blood cell images using
474 convolutional neural networks. *Comput. Methods Programs Biomed.* **180**, 105020, [10.1016/j.cmpb.2019.105020](https://doi.org/10.1016/j.cmpb.2019.105020) (2019).
- 475 **76.** Dietler, N. *et al.* A convolutional neural network segments yeast microscopy images with high accuracy. *Nat. Commun.* **11**,
476 [10.1038/s41467-020-19557-4](https://doi.org/10.1038/s41467-020-19557-4) (2020).
- 477 **77.** Data and Software — epfl.ch. <https://www.epfl.ch/labs/lpbs/data-and-software/>. [Accessed 07-08-2024].

478 **Acknowledgements**

479 We would like to express our gratitude for the authors of the MedMNIST⁸, which served as the baseline for our study, and for
480 the shared source code that we referenced to develop our own code.

481 **Author contributions statement**

482 M.A. conducted data collection, cleaning and pre-processing steps. Z.K. performed the evaluation tests for binary and
483 multi-class task for each network and datasets. All authors wrote and reviewed the manuscript.

484 **Competing interests**

485 The authors state that they have no conflicting interests.