scientific data

Check for updates

OPEN

DATA DESCRIPTOR

MedSegBench: A comprehensive benchmark for medical image segmentation in diverse data modalities

Zeki Kuş ✉ & Musa Aydin

MedSegBench is a comprehensive benchmark designed to evaluate deep learning models for medical image segmentation across a wide range of modalities. It covers a wide range of modalities, including 35 datasets with over 60,000 images from ultrasound, MRI, and X-ray. The benchmark addresses challenges in medical imaging by providing standardized datasets with train/validation/test splits, considering variability in image quality and dataset imbalances. The benchmark supports binary and multi-class segmentation tasks with up to 19 classes and uses the U-Net architecture with various encoder/decoder networks such as ResNets, EfficientNet, and DenseNet for evaluations. MedSegBench is a valuable resource for developing robust and flexible segmentation algorithms and allows for fair comparisons across different models, promoting the development of universal models for medical tasks. It is the most comprehensive study among medical segmentation datasets. 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 diseases[1]. 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 networks[2]. 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 datasets[3]. 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.

Limited benchmark studies in the literature focus on medical imaging, with most concentrating on medical image classification problems[4–8]. 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 Decathlon[6] supports creating and benchmarking semantic segmentation algorithms. It includes 2633 3D images from ten 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 supports 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,

Fatih Sultan Mehmet Vakif University, Computer Engineering, İstanbul, 34445, Türkiye. ✉e-mail: zkus@fsm.edu.tr

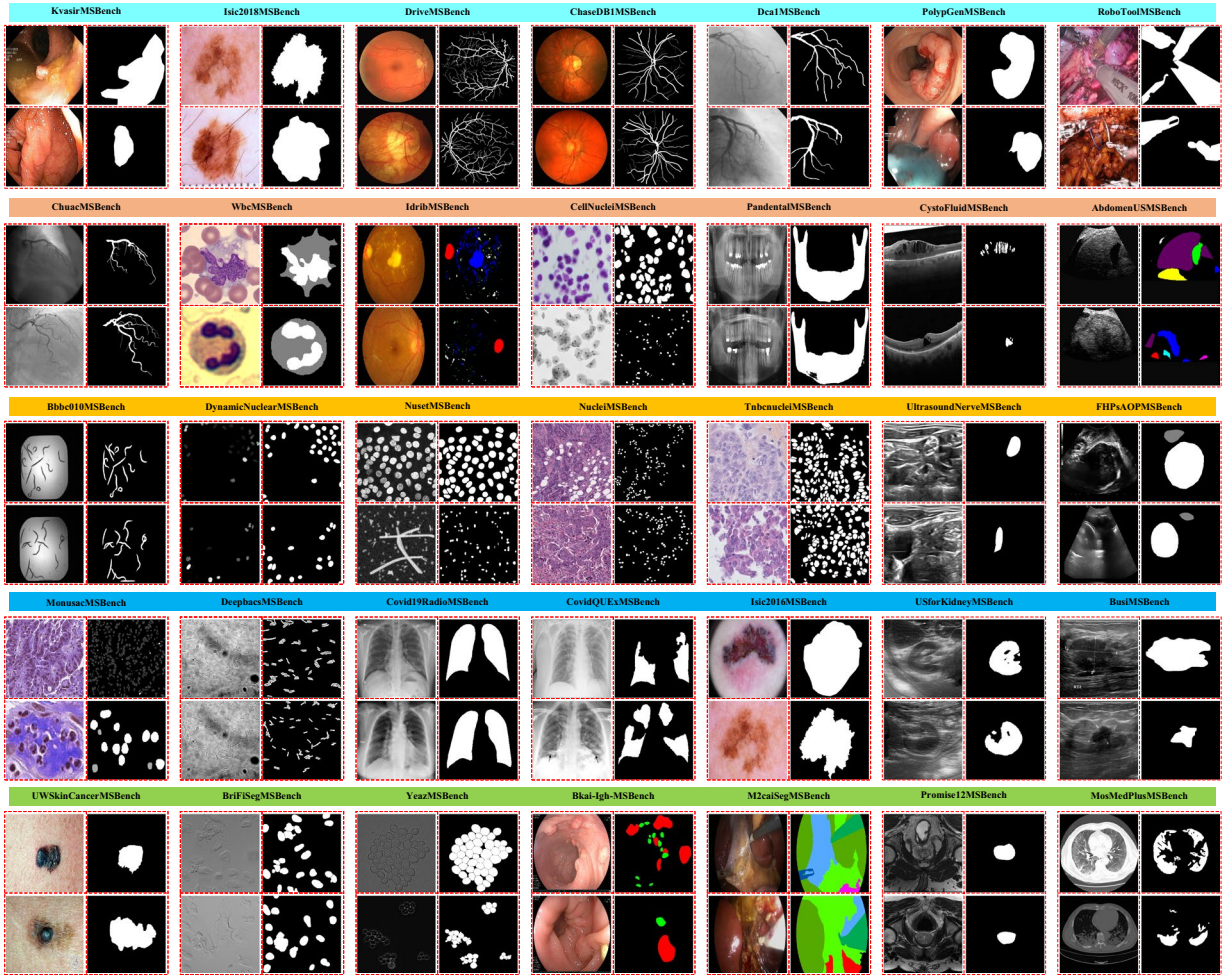

Fig. 1 Visual overview of the 35 datasets included in MedSegBench. Each dataset is represented by two sample images, showcasing the diversity of medical imaging modalities and segmentation tasks covered in this benchmark. The datasets span various anatomical regions and pathologies, including abdominal ultrasound, cell microscopy, chest X-rays, dermoscopy, endoscopy, fundus imaging, MRI, CT scans, and more.

This study introduces a comprehensive benchmark dataset for medical image segmentation (Fig. 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 binary and multi-class problems, with some datasets featuring up to 19 classes. This benchmark offers several powerful advantages as a robust and versatile tool for the research community:

- **Diversity of modalities:** The benchmark includes datasets from various imaging modalities such as Ultrasound, MRI, X-Ray, OCT, Dermoscopy, Endoscopy, and various types of microscopy.
- **Task complexity:** It covers binary and multi-class segmentation tasks with up to 19 classes.
- **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.
- **Data split:** All datasets follow a standard train/validation/test split, which is crucial for properly evaluating machine learning models.
- **Standardization:** All datasets are standardized to enhance comparability and ease of use. Samples across all datasets have been resized to three standard resolutions - 128, 256, and 512 pixels - and stored in a uniform format.
- **Application areas:** The datasets cover various medical applications, including cancer detection, COVID-19 diagnosis, cell and nuclei segmentation, and organ segmentation.

We have evaluated each dataset on state-of-the-art segmentation model (U-Net[9]) with different encoder/decoder network types (ResNet-18, ResNet-50, Efficient-Net, MobileNet-v2, DenseNet-121, Mix Vision Transformer)[10]. Each experiment is performed 3 times, and average results are reported.

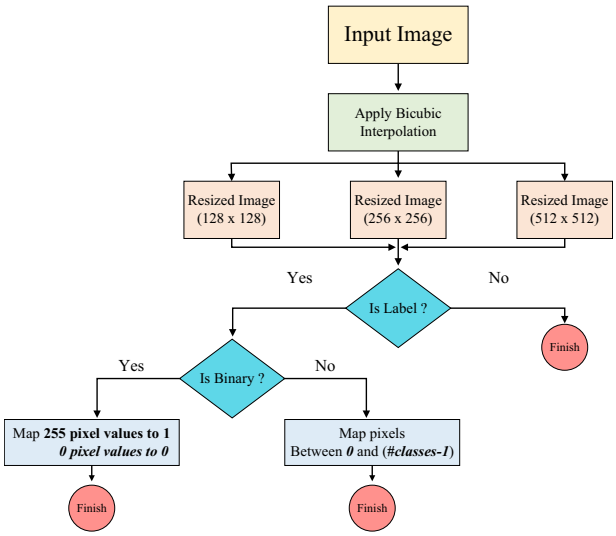

Fig. 2 Flowchart of an image preprocessing pipeline: An input image is resized using bicubic interpolation to dimensions of 128 × 128, 256 × 256, or 512 × 512 pixels. Following this are two decision points: Is Label? and Is Binary?. If the image is not a label, it proceeds to finish. If it is a binary label, it maps pixel values (255 to 1, 0 to 0). If it's a label and not binary, it maps pixels between 0 and (#classes-1) before finishing.

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

Methods

Data preparation. *Dataset selection and standardization.* The MedSegBench dataset[11] comprises 35 distinct 2D medical image segmentation datasets, some of which are extracted from 3D slices. These datasets cover various data modalities such as Ultrasound, OCT, Chest X-ray, MR, and more. The original datasets differ in scales, segmentation tasks (binary/multi-class), classes, imaging modalities, and annotation styles. Hence, we have selected a standardized format and performed pre-processing to ensure a consistent format across all datasets.

Image resizing and label mapping. Numerous medical image segmentation datasets are available in the literature, each presenting various challenges, including variations in annotations, image sizes, and file formats. Additionally, many of these datasets lack officially shared train/test/validation splits, making it challenging to compare different methods fairly. To address these issues, we performed pre-processing steps. All image and label pairs are resized to 128 × 128, 256 × 256, and 512 × 512 pixels using the bicubic interpolation method. Bicubic interpolation produces smoother images using a 4 × 4 pixel neighborhood, preserving details crucial for medical imaging[2]. It balances computational efficiency and image quality, making it ideal for large datasets like MedSegBench. Although we used 512 × 512 sized images in our experiments, we have made the 128 × 128 and 256 × 256 sized versions publicly available for researchers with limited GPU memory. Also, we have applied a mapping to labels; pixels with values of 0 and 255 are mapped to 0 and 1 for binary segmentation tasks, and for multi-class segmentation tasks, pixels are mapped to integer values between 0 and (#Classes - 1). No additional augmentation or pre-processing steps are applied to the images and labels. Figure 2 has presented preprocessing steps.

Train/Validation/Test Splits. We have followed three different scenarios based on MedMNIST v2[8] to create train/test/validation splits, using a seed value of 42 for consistency and reproducibility: (1) Utilizing the source train/test/validation splits if published by the authors; (2) Using the source validation set as the test set and splitting the source training set into 90% training and 10% validation (9:1 ratio) if the source training and validation splits are published by the authors[12]; (3) Randomly splitting the dataset into 70% training, 10% validation, and 20% test sets if no public train/test/validation splits are available (7:1:2 ratio)[13,14].

Data storage and licensing. Most of these datasets are publicly published under Creative Commons Licenses, some of which are CC-BY-NC, CC-BY-SA, and CC-BY-NC-SA, permitting the redistribution of datasets. In addition, other datasets whose license status is not disclosed are shared publicly for educational purposes, and re-distribution permission has been obtained from the authors by email. We have published datasets in MedSegBench under Creative Commons Licences, and source codes have been published under Apache License 2.0.

Table 1 presents the summary information for all MedSegBench datasets. In addition, Table 2 shows the data-modality-based overview for MedSegBench datasets. Furthermore, Table 3 provides an overview of various

Dataset Name[source]	Modality	Pathology/Organ Studied	Binary or Multi-class (# Classes)	# Images	# Train/Val/Test
AbdomenUSMSBench[15,16]	Ultrasound	Gallbladder, Kidney, Liver, Spleen, Vessel	Multi-class (9)	926	569/64/293
Bbbc010MSBench[17,18]	Microscopy	Caenorhabditis elegans	Binary	100	70/10/20
Bkai-Igh-MSBench[19–21]	Endoscopy	Colon polyps	Multi-class (3)	1,000	700/100/200
BriFiSegMSBench[22,23]	Microscopy	Lung, Cervix, Breast, Eye	Binary	1,360	1005/115/240
BusiMSBench[24,25]	Ultrasound	Breast	Binary	647	452/64/131
CellNucleiMSBench[26,27]	Nuclei	Nuclei	Binary	670	469/67/134
ChaseDB1MSBench[28]	Fundus	Eye (Retinal vessels)	Binary	28	19/2/7
ChuacMSBench[29]	Fundus	Eye (Retinal vessels)	Binary	30	21/3/6
Covid19RadioMSBench[30–32]	Chest X-Ray	Lung	Binary	21,165	14,814/2,115/4,236
CovidQUExMSBench[33,34]	Chest X-Ray	Lung	Binary	2,913	1,864/466/583
CystoFluidMSBench[37–39]	OCT	Eye (Cystoid macular edema)	Binary	1,006	703/101/202
Dca1MSBench[40,41]	Fundus	Eye (Retinal vessels)	Binary	134	93/13/28
DeepbacsMSBench[42,43]	Microscopy	Bacterial cells	Binary	34	17/2/15
DriveMSBench[44,45]	Fundus	Eye (Retinal vessels)	Binary	40	18/2/20
DynamicNuclearMSBench[46,47]	Nuclear Cell	Nuclear Cells	Binary	7,084	4,950/1,417/717
FHPsAOPMSBench[48,49]	Ultrasound	Fetal head, pubic symphysis	Multi-class (3)	4,000	2,800/400/800
IdribMSBench[50,51]	Fundus	Eye (Optic discs)	Binary	80	47/6/27
Isic2016MSBench[52,53]	Dermoscopy	Skin (Lesions)	Binary	1,279	810/90/379
Isic2018MSBench[54–56]	Dermoscopy	Skin (Lesions)	Binary	3,694	2,594/100/1,000
KvasirMSBench[57,58]	Endoscopy	Gastrointestinal polyps	Binary	1,000	700/100/200
M2caiSegMSBench[59,60]	Endoscopy	Surgical tools and abdominal tissues	Multi-class (19)	614	245/307/62
MonusacMSBench[61,62]	Pathology	Lung, Prostate, Kidney, and Breast	Multi-class (6)	310	188/21/101
MosMedPlusMSBench[35,36]	CT	Lung	Binary	2,729	1,910/272/547
NucleiMSBench[63]	Pathology	Cell Nuclei	Binary	141	98/14/29
NusetMSBench[64,65]	Nuclear Cell	Nuclear cells	Binary	3,408	2,385/340/683
PandentalMSBench[66,67]	X-Ray	Mandible	Binary	116	81/11/24
PolypGenMSBench[68,69]	Endoscopy	Colon polyps	Binary	1,412	984/140/288
Promise12MSBench[70,71]	MRI	Prostate	Binary	1,473	1,031/147/295
RoboToolMSBench[37]	Endoscopy	Surgical tools	Binary	500	350/50/100
TnbcnucleiMSBench[72,73]	Pathology	Nuclei in histopathology images	Binary	50	35/5/10
UltrasoundNerveMSBench[74]	Ultrasound	Neck (Brachial Plexus Nerves)	Binary	2,323	1,651/223/449
USforKidneyMSBench[75,76]	Ultrasound	Kidney	Binary	4,586	3,210/458/918
UWSkinCancerMSBench[77]	Dermoscopy	Skin (Cancer)	Binary	206	143/19/44
WbcMSBench[78,79]	Microscopy	White Blood Cell	Multi-class (3)	400	280/40/80
YeazMSBench[80,81]	Microscopy	Yeast Cells	Binary	707	360/96/251

Table 1. An overview of the 35 datasets included in MedSegBench for medical image segmentation. For each dataset, the table lists the dataset name with source references, imaging modality, the pathology or organ studied, whether the segmentation task is binary or multi-class (along with the number of classes), the total number of images, and the distribution of images into training, validation, and test sets. The datasets cover various imaging modalities such as Ultrasound, MRI, X-ray, OCT, and others, covering a wide range of medical applications and challenges.

datasets, detailing their sub-categories and the number of samples for training, validation, and testing. In the following sections, we will describe the details of each dataset.

Details. *AbdomenUSMSBench.* The AbdomenUSMSBench created from AbdomenUS[15,16] consists of 926 ultrasound images of the abdominal region, each with a resolution of 449 × 464 pixels. This dataset is designed for multi-class segmentation tasks and includes nine distinct classes. We have used the official train and test splits, and the train set is split into a training and 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 between 0 and (#Classes - 1).

Bbbc010MSBench. The Bbbc010MSBench dataset derived from Bbbc010[17,18], 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

Modality	Number of Images	Total Number of Datasets
Computed Tomography	2,729	1
Dermoscopy	5,179	3
Endoscopy	4,526	5
Fundus	312	5
Magnetic Resonance Imaging	1,473	1
Microscopy	2,281	5
Nuclear Cell	10,492	2
Nuclei	670	1
Optical Coherence Tomography	1,006	1
Pathology	501	3
Ultrasound	12,482	5
X-Ray	24,194	3

Table 2. Overview of the medical imaging modalities represented in the MedSegBench benchmark datasets. For each modality, the table lists the total number of images and datasets included in MedSegBench.

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

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.

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 dataset[19–21] 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[22,23], 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[24,25] and contains 647 ultrasound images with an average resolution of 500 × 500 pixels. This dataset is designed for binary segmentation tasks, categorizing data into benign and malignant classes. 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 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

CellNucleiMSBench. The CellNucleiMSBench comes from the 2018 Data Science Bowl[26,27] 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 is split into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

ChaseDB1MSBench. ChaseDB1MSBench is based on the CHASE_DB1 dataset[28], 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: train/val/test, in a 7:1:2 ratio. The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

ChuacMSBench. The ChuacMSBench, derived from the CHUAC dataset[29], includes 28 fundus images with 189 × 189 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. The samples are resized to 1 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

Covid19RadioMSBench. The COVID-19 Radiography Database[30–32] is the source of the Covid19RadioMSBench dataset, which consists of 21,165 chest X-ray images, each with a resolution of 299 × 299 pixels. This dataset is designed for binary 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 collaborative effort of researchers from Qatar University, the University of Dhaka, and partners from Pakistan and Malaysia, working alongside medical professionals. It includes chest X-ray images for COVID-19 positive cases and Normal and Viral Pneumonia images. The authors have also categorized the images into four groups: COVID, Lung_Opacity, Normal, and Viral Pneumonia. We provide these category-based images and their corresponding annotations in.npz file format (see Table 3). The samples are resized to 1 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

CovidQUExMSBench. The CovidQUExMSBench, based on the COVID-QU-Ex Dataset[33,34], consists of 2,913 chest X-ray images, each with a resolution of 256 × 256 pixels. This dataset is specifically designed for binary segmentation tasks. We use only infection segmentation samples. The source dataset is split into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to 1 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

MosMedPlusMSBench. The MosMedPlusMSBench, based on the MosMedDataPlus[35,36] dataset, comprises 2,729 Covid-19 CT images, each sized 512 × 512 pixels. This dataset is designed for binary segmentation tasks. We split source data into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

CystoFluidMSBench. The CystoFluidMSBench is based on Intraretinal Cystoid Fluid dataset[37–39], comprises 1,006 OCT (Optical Coherence Tomography) images, most of which are sized at 512 × 512 pixels. This dataset is designed for binary segmentation tasks. The images are carefully chosen by medical experts at Liaquat University of Medical and Health Sciences (LUMHS) Jamshoro, who are trained to identify Cystoid Macular Edema (CME) and its progression, providing a confirmatory diagnosis of CME. The source dataset is split into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

Dca1MSBench. The Dca1MSBench is derived from the DCA1 dataset[40,41] and contains 134 fundus images, each with a resolution of 300 × 300 pixels. The Cardiology Department of the Mexican Social Security Institute, UMAE T1-LeÃ³n, provides the images. This dataset is specifically created for binary segmentation tasks. The dataset is split into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to 1 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

DeepbacsMSBench. The DeepbacsMSBench, based on the DeepBacs dataset[42,43], consists of 34 samples of fundus images, each with a size of 1024 × 1024 pixels. It is designed for binary segmentation tasks. We use official train/validation/test splits published officially by authors. The samples are resized to 1 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

DriveMSBench. The DriveMSBench dataset, based on the DRIVE dataset[44,45], includes 40 fundus images, each with dimensions of 565 × 584 pixels. The images are obtained from a diabetic retinopathy screening program in the Netherlands. It is designed for binary segmentation and uses official splits for training, validation, and testing. The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

DynamicNuclearMSBench. The DynamicNuclearMSBench, created from the DynamicNuclearNet Segmentation dataset[46,47], consists of 7084 samples of nuclear cell images, each 128 × 128 pixels in size. This dataset is utilized for a binary segmentation task. Training, validation, and test splits that are officially published are used. The samples are resized to 1 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

FHPsAOPMSBench. The FHPsAOPMSBench dataset is based on a prior dataset[48,49] and comprises 4,000 ultrasound images, each with a resolution of 256 × 256 pixels. This dataset is designed for a multi-class segmentation task, including 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 1 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and (#Classes - 1).

IdribMSBench. The IdribMSBench is based on the Indian Diabetic Retinopathy Image Dataset[50,51] and includes 80 high-resolution fundus images (4288 × 2848 pixels) for a binary segmentation task. We use official train/validation/test splits published officially by authors. The authors have also categorized the labels into four categories: Microaneurysms, hemorrhages, Hard Exudates, and Optic Discs. These category-based labels and annotations are provided in a npz file (see Table 3). The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

Isic2016MSBench. The Isic2016MSBench is derived from the ISIC 2016 Challenge[52,53], which consisted of 1,279 dermoscopy samples of varying sizes designed for binary segmentation tasks. We use official training, validation, and test splits published by authors. The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

Isic2018MSBench. The Isic2018MSBench is derived from the ISIC 2018 Challenge[54–56], which consisted of 3,694 dermoscopy samples of varying sizes designed for binary segmentation tasks. We use official training, validation, and test splits published by authors. The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

KvasirMSBench. The KvasirMSBench, derived from the Kvasir-SEG dataset[57,58], consists of 1,000 endoscopy images with resolutions ranging from 332 × 487 to 1920 × 1072 pixels. The dataset includes images of gastrointestinal polyps and their segmentation masks, which an experienced gastroenterologist has annotated and verified. It is structured for a binary classification task. The source dataset is divided into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

M2caiSegMSBench. M2caiSegMSBench is based on a prior dataset[59,60] comprising 614 pathology samples and designed for multi-class segmentation tasks, including 19 distinct classes. The images within this dataset exhibit variable dimensions, and we use official train/validation/test splits. The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and (#Classes - 1).

MonusacMSBench. MonusacMSBench is based on the MoNuSAC challenge[61,62]. It consists of 310 samples and is designed for multi-class segmentation with 6 classes. The images in this dataset are H&E stained digitized tissue images from several patients acquired at multiple hospitals using a standard 40x scanner magnification. The annotations are provided by expert pathologists. We use the officially published train/validation/test splits from the challenge. The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and (#Classes - 1).

NucleiMSBench. The NucleiMSBench is based on a prior dataset[63], consisting of 141 pathology samples, each with an image size of 2000 × 2000 pixels. This source dataset is 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 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

NusetMSBench. The NusetMSBench, derived from the NuSet dataset[64,65], contains 3,408 pathology samples designed for binary segmentation problems. We split the source dataset into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to 1 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

PandentalMSBench. The PandentalMSBench is created from the Panoramic Dental X-rays dataset[66,67] and contains 116 X-ray samples of varying sizes. It is specifically intended for binary segmentation tasks. The dataset comprises anonymized and deidentified panoramic dental X-rays of 116 patients taken at Noor Medical Imaging Center in Qom, Iran. The source dataset is divided into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to 1 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

PolypGenMSBench. The PolypGenMSBench is based on a prior endoscopy dataset[68,69] consisting of 1,412 endoscopy samples, each with an image size of 1920 × 1080 pixels. It is designed for binary segmentation tasks. It includes colonoscopy video frames captured from a diverse patient population at six different centers in Egypt, France, Italy, Norway, and the United Kingdom. We provide these images, and annotations are captured from these centers in a npz file. The source dataset is divided into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

Promise12MSBench. The Promise12MSBench, derived on the PROMISE12 dataset[70,71], contains 1,473 MR samples, each with an image size of 512 × 512 pixels. It is designed for binary classification. We split the source dataset into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to 1 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

RoboToolMSBench. The RoboToolMSBench, based on the RoboTool dataset[37], consisting of 500 samples, is designed for binary segmentation tasks. The source dataset is divided into three parts: train/val/test, in a 7:1:2 ratio. The samples are resized to 1 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

TnbcnucleiMSBench. The TnbcnucleiMSBench is based on a prior dataset[72,73], consisting of 50 pathology samples, each with an image size of 512 × 512 × pixels. This dataset is based on the merging of two different datasets: the first dataset, generated at the Curie Institute, consists of annotated H&E stained histology images at 40× magnification, and the second dataset, provided by the Indian Institute of Technology Guwahati, also consists of annotated H&E stained histology images captured at 40× magnification. It 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 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

UltrasoundNerveMSBench. The UltrasoundNerveMSBench, derived from prior dataset[74], contains 2,323 ultrasound samples, each with an image size of 580 × 420 pixels and designed for binary segmentation tasks. The primary task in this 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 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

USforKidneyMSBench. The USforKidneyMSBench is derived from the CT2USforKidneySeg dataset[75,76], 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 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

UWSkinCancerMSBench. The UWSkinCancerMSBench is based on the Skin Cancer Detection dataset[77], 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 Melenoma and Not-Melenoma. These category-based labels and annotations are provided in a.npz file (see Table 3). The samples are resized to 3 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and 1.

WbcMSBench. The WbcMSBench, based on prior datasets[78,79], 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 × 512 × 512 pixels, and the labels are mapped to integer values between 0 and (#Classes - 1).

YeazMSBench. The YeazMSBench, derived from the YeaZ dataset[80,81], 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 × 512 × 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[11]. 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 MedMNISTv2[8] 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 explained 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 includes train, validation and test labels with $N \times W \times H$ shape. **{train,val,test}_images_{classno} and {train,val,test}_label_{classno}:** These hold 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. ResNet18 is chosen over ResNet34 primarily due to its lower computational complexity and faster training times, which are advantageous when working with large datasets or limited computational resources. Despite being shallower, ResNet18 provides a good balance between depth and efficiency, making it suitable for capturing essential features in medical images, especially when combined with the U-Net architecture that enhances spatial resolution through its encoder-decoder structure[82,83]. ResNet50, with its deeper architecture, offers more detailed feature extraction capabilities, allowing it to handle more complex segmentation tasks. DenseNet121 is included for its efficient use of parameters and its dense connectivity, which mitigates the vanishing gradient problem and is beneficial for capturing fine details in medical images[84]. We have also selected EfficientNet and MobileNetv2 because they are lightweight models offering 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 framework[10]. 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 of 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 corresponding to the best validation IOU are recorded for each network configuration. Further details regarding the model implementation, training, and evaluation steps are available in our code repository.

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

Results

The average PREC and REC results obtained from three different runs are shown in Table 4, and average F1 and IOU scores are reported in Table 5 for each individual dataset. Also, Table 5 shows the average results for each baseline method. Additionally, we have provided detailed performance metrics for each seed and model across all datasets, alongside image-wise performance metrics to assess robustness and reliability. All related data, including model weights for each model with three seeds, are available on our Zenodo page[85] and Github repository (see Code availability section).

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

Table 5 provides a comprehensive evaluation of six encoder networks across various datasets, using F1-score and Intersection over Union (IOU) as performance metrics. DenseNet-121 consistently performs well, frequently yielding the top F1 and IOU metrics scores across numerous datasets. For example, in the Bbbc010MSB and CellNucleiMSB datasets, DenseNet-121 records the highest F1-scores of 0.920 and 0.907, respectively, and similarly high IOU scores, indicating its robustness in handling diverse data types. Efficient-Net also shows significant performance, particularly in datasets like Isic2016MSB and USforKidneyMSB, where it achieves the highest F1-scores of 0.903 and 0.981, respectively. This indicates that Efficient-Net is particularly effective in scenarios requiring high precision and recall, as shown in its F1 scores. ResNet-50 performs best with an F1-score of 0.931 and an IOU of 0.870 for the DeepbacsMSB. Additionally, it has also attained the highest F1-score of 0.786 and an IOU of 0.648 in the DriveMSB dataset. For the FHPsAOPMSB dataset, ResNet-18 has achieved

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**	**0.589**	**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	0.905	**0.906**	—
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**
TnbcnucleiMSB	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 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. Results are presented for each dataset, with the highest scores for precision and recall highlighted. A dash (—) indicates that the network is not evaluated for that particular dataset due to input channel constraints.

the highest F1-score of 0.961 and an IOU of 0.929. While Mix Vision Transformer does not frequently perform as well as DenseNet-121, it shows competitive performance in specific datasets such as UWSkinCancerMSB, achieving the second-highest F1 Score of 0.881. This indicates its potential in specialized applications, particularly in medical imaging contexts. Overall, DenseNet-121 is the most robust and effective network, frequently outperforming other networks in yielding high F1-scores and IOU values. Table 5 has also demonstrated how the F1-score provides clearer insights than Precision or Recall alone. The CellNucleiMSBench consists of 670 images for nuclei segmentation tasks, where the number of nuclei (positive class) is significantly lower compared to the background (negative class). For example, DenseNet-121 achieved a Precision of 0.927 and a Recall of 0.898 on this dataset, resulting in an F1-score of 0.907. This high F1-score reflects a balanced performance, ensuring the model is precise and sensitive in nuclei detection. In another example, the Bbbc010MSBench with 100 microscopy images designed for binary segmentation, the distribution between the classes can be uneven, especially in identifying specific cellular structures. DenseNet-121 has achieved a Precision of 0.922 and a Recall of 0.920, resulting in an F1-score of 0.920. This demonstrates that the model effectively balances precision and recall. Lastly, Isic2018MSBench involves segmenting skin lesions, where the area covered by lesions (positive class) can vary widely compared to healthy skin (negative class). EfficientNet attained a Precision of 0.857 and a Recall of 0.923, resulting in an F1-score of 0.868. This indicates that the model maintains a strong balance between accurately identifying lesions and minimizing missed detections.

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**
TnbcnucleiMSB	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 5. The average F1-score and IOU results for six 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. Results are presented for each dataset, with the highest scores for F1 and IOU highlighted. A dash (—) indicates that the network was not evaluated for that particular dataset due to input channel constraints.!

Table 6 shows the mean and standard deviation of the performance metrics for six different encoder networks. Efficient-Net (EN) and DenseNet-121 (DN-121) demonstrate the highest F1 scores, both achieving a value of 0.772. This indicates that these models have a balanced performance in terms of precision and recall. DenseNet-121 also obtains the highest precision at 0.848 with a standard deviation of ± 0.011, indicating reliable precision across datasets and effectively minimizing false positives. On the other hand, Efficient-Net leads in recall with a score of 0.788 with a standard deviation of ± 0.017, indicating its strength in capturing true positives. Additionally, DenseNet-121 reaches the highest IOU of 0.702 with a standard deviation of ± 0.010, indicating stable performance. This is closely followed by Efficient-Net, which achieves an IOU of 0.700. It suggests that these two models provide the most accurate predictions. Overall, DenseNet-121 and Efficient-Net reached similar high-performance metrics, with both models performing well in F1 score, precision, recall, and IOU. However, DenseNet-121's complex architecture causes higher computational demands, whereas Efficient-Net provides a more efficient design, making it suitable for resource-constrained applications. When analyzing the performance of DenseNet-121 (showing the best average performance on 3 of the 4 measures) across different datasets, no specific characteristic is observed. DenseNet-121 performs well across diverse datasets, excelling in handling class imbalance (CellNucleiMSB, Isic2016MSB, Isic2018MSB) and complex multi-class segmentation tasks (FHPsAOPMSB, WbcMSB). Its architecture, featuring dense connections, enables effective learning from minority classes and intricate feature hierarchies. It achieves good results for datasets with either high (CovidQUExMSB, DynamicNuclearMSB, Isic2018MSB) or low sample sizes (ChaseDB1MSB, ChuacMSB,

Methods	F1	PREC	REC	IOU
RN-18	0.762 ± 0.008	0.834 ± 0.011	0.774 ± 0.014	0.689 ± 0.008
RN-50	0.759 ± 0.010	0.833 ± 0.010	0.772 ± 0.017	0.686 ± 0.010
EN	**0.772** ± 0.011	0.832 ± 0.012	**0.788** ± 0.017	0.700 ± 0.012
MN-v2	0.762 ± 0.009	0.834 ± 0.010	0.769 ± 0.013	0.689 ± 0.009
DN-121	**0.772** ± 0.009	**0.848** ± 0.011	0.781 ± 0.014	**0.702** ± 0.010
MVT	0.663 ± 0.016	0.760 ± 0.019	0.696 ± 0.021	0.585 ± 0.017

Table 6. Average and standard deviation of performance metrics for six different encoder networks across all datasets in MedSegBench. RN-18: ResNet-18; RN-50: ResNet-50; EN: Efficient-Net; MN-v2: Mobile-Net-v2; DN-121: DenseNet-121; MVT: Mix Vision Transformer. The highest score for each metric is highlighted in bold.

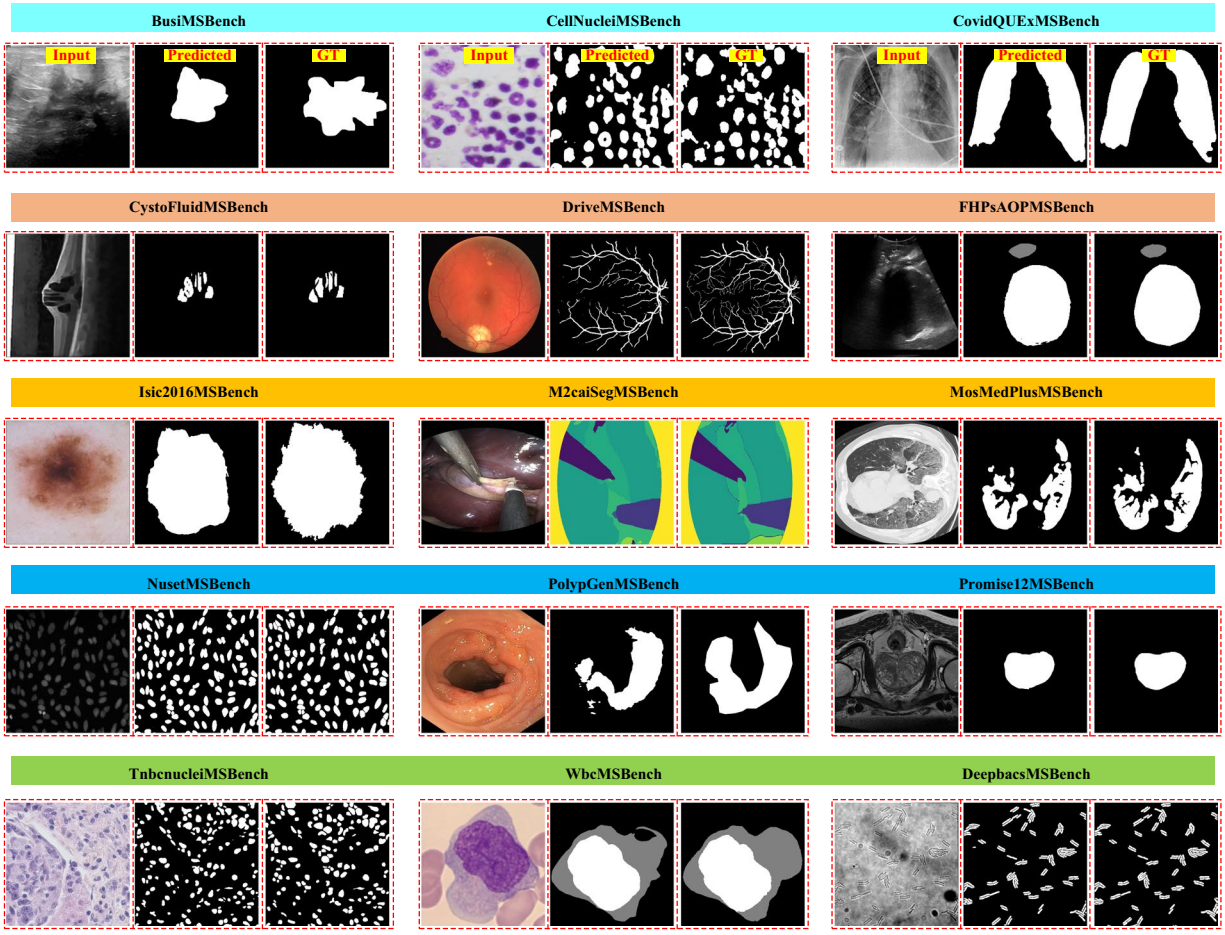

Fig. 3 Segmentation results for selected datasets across different modalities, showcasing the performance of the best models as reported in Table 5. Each row includes three datasets, with columns representing the input image, the predicted segmentation by the selected model, and the ground truth segmentation.

YeazMSB). Additionally, it performs well with datasets from different data modalities (BriFiSegMSB, KvasirMSB, Promise12MSB). While it generally outperforms other methods in multi-class problems, its performance is comparatively lower than in binary segmentation tasks. Overall, DenseNet-121 demonstrates versatility across various imaging modalities and dataset sizes, with consistent performance indicated by low standard deviations in metrics. These characteristics make it a reliable choice for diverse medical imaging tasks.

Datasets with classes that look similar present challenges for segmentation models because it's hard to tell them apart. In our study, models have performed consistently across datasets like ultrasound images of benign and malignant lesions, WbcMSBench with cytoplasm and nucleus segmentation, and fundus images in the ChaseDB1MSB and DriveMSB datasets with similar retinal vessel patterns (Tables 4 and 5). This shows that our benchmark effectively evaluates models on tasks with structurally similar classes. However, the ChuacMSB dataset has performed less than other retinal vessel segmentation tasks. This might be due to its different imaging techniqueâ€"coronary angiographyâ€"which introduces unique visual features. The models also showed

consistent results on dermoscopy datasets (Isic2016MSB, Isic2018MSB, and UWSkinCancerMSB) and COVID-19 CT images (CovidQUExMSB and MosMedPlusMSB), highlighting their robustness with visually similar classes. These results show the models' strengths and areas for improvement when dealing with classes that resemble each other.

We also visually compared the results. Figure 3 illustrates the segmentation results for selected datasets. We have selected one dataset from each data modality (two for some modalities due to binary/multi-class tasks) to maintain diversity. Each row displays three datasets, with the columns showing the inputs, predicted images by the best model according to IOU and the corresponding ground truth labels. The figure shows that the top models have achieved strong segmentation results compared to the ground truth labels. However, there are some misclassified pixels for both binary and multi-class segmentation tasks. Multi-class segmentation presents a significant challenge for segmentation models, even the top-performing ones. This comprehensive visualization demonstrates the effectiveness and versatility of the models across various imaging modalities and medical conditions.

In conclusion, MedSegBench[11] represents a significant advancement in medical image segmentation by providing a comprehensive benchmark that spans a wide array of imaging modalities and segmentation tasks. With 35 datasets and over 60,000 images, it offers a robust benchmark for evaluating the performance of deep learning models, particularly highlighting the effectiveness of DenseNet-121 and Efficient-Net architectures. Despite its strengths, the benchmark primarily focuses on 2D image segmentation, which may not fully address the complexities of 3D medical imaging. Future research could expand this work by incorporating more 3D datasets and exploring the potential of transformer-based models to enhance segmentation accuracy further.

Usage Notes

This dataset was created to compare different models fairly over various segmentation models from different data modalities and create universal models. It is not suitable for clinical or medical use.

Code availability

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

Received: 23 August 2024; Accepted: 19 November 2024;
Published online: 25 November 2024

References

1. Han, K. *et al*. Deep semi-supervised learning for medical image segmentation: A review. *Expert Systems with Applications* **245**, 123052, https://doi.org/10.1016/j.eswa.2023.123052 (2024).
2. Ma, J. *et al*. Segment anything in medical images. *Nature Communications* **15**, https://doi.org/10.1038/s41467-024-44824-z (2024).
3. Carriero, A., Groenhoff, L., Vologina, E., Basile, P. & Albera, M. Deep learning in breast cancer imaging: State of the art and recent advancements in early 2024. *Diagnostics* **14**, 848, https://doi.org/10.3390/diagnostics14080848 (2024).
4. Drelie Gelasca, E., Obara, B., Fedorov, D., Kvilekval, K. & Manjunath, B. A biosegmentation benchmark for evaluation of bioimage analysis methods. *BMC Bioinformatics* **10**, https://doi.org/10.1186/1471-2105-10-368 (2009).
5. Rebuffi, S.-A., *et al*. (eds.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017).
6. Simpson, A. L. *et al*. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *CoRR* abs/1902.**09063** (2019).
7. Yang, J., Shi, R. & Ni, B. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, https://doi.org/10.1109/isbi48211.2021.9434062p (IEEE, 2021).
8. Yang, J. *et al*. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data* **10**, https://doi.org/10.1038/s41597-022-01721-8 (2023).
9. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, 234–241 (Springer International Publishing, 2015).
10. Iakubovskii, P. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch (2019).
11. Aydin, M. & KUŞ, Z. Medsegbench: A comprehensive benchmark for medical image segmentation in diverse data modalities, https://doi.org/10.5281/ZENODO.13359660 (2024).
12. Xu, Y. & Goodacre, R. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing* **2**, 249–262, https://doi.org/10.1007/s41664-018-0068-2 (2018).
13. Ma, J. J. *et al*. Diagnostic image quality assessment and classification in medical imaging: Opportunities and challenges. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 337–340, https://doi.org/10.1109/ISBI45749.2020.9098735 (2020).
14. Singh, P. *et al*. Shifting to machine supervision: annotation-efficient semi and self-supervised learning for automatic medical image segmentation and classification. *Scientific Reports* **14**, https://doi.org/10.1038/s41598-024-61822-9 (2024).
15. Vitale, S., Orlando, J. I., Iarussi, E. & Larrabide, I. Improving realism in patient-specific abdominal ultrasound simulation using cyclegans. *International Journal of Computer Assisted Radiology and Surgery* **15**, 183–192, https://doi.org/10.1007/s11548-019-02046-5 (2019).
16. Orlando, J. I. Us simulation & segmentation (2020).
17. Ljosa, V., Sokolnicki, K. L. & Carpenter, A. E. Annotated high-throughput microscopy image sets for validation. *Nature Methods* **9**, 637–637, https://doi.org/10.1038/nmeth.2083 (2012).
18. Broad Bioimage Benchmark Collection — bbbc.broadinstitute.org. https://bbbc.broadinstitute.org/BBBC010. [Accessed 06-08-2024].
19. Ngoc Lan, P. *et al*. *NeoUNet: Towards Accurate Colon Polyp Segmentation and Neoplasm Detection*, 15–28 (Springer International Publishing, 2021).
20. An, N. S. *et al*. Blazeneo: Blazing fast polyp segmentation and neoplasm detection. *IEEE Access* **10**, 43669–43684, https://doi.org/10.1109/access.2022.3168693 (2022).
21. Duc, N. T., Oanh, N. T., Thuy, N. T., Triet, T. M. & Dinh, V. S. Colonformer: An efficient transformer based method for colon polyp segmentation. *IEEE Access* **10**, 80575–80586, https://doi.org/10.1109/access.2022.3195241 (2022).
22. Mathieu, G., Annika, L. & Bachir, E. D. Brifiseg: a deep learning-based method for semantic and instance segmentation of nuclei in brightfield images, https://doi.org/10.48550/ARXIV.2211.03072 (2022).

23. Gendarme, M. & Debs, B. E. Brifiseg datasets, https://doi.org/10.5281/ZENODO.7195636 (2022).
24. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data in Brief* **28**, 104863, https://doi.org/10.1016/j.dib.2019.104863 (2020).
25. Breast Ultrasound Images Dataset — kaggle.com. https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset. [Accessed 06-08-2024].
26. Caicedo, J. C. *et al*. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature Methods* **16**, 1247–1253, https://doi.org/10.1038/s41592-019-0612-7 (2019).
27. 2018 Data Science Bowl — kaggle.com. https://www.kaggle.com/competitions/data-science-bowl-2018/data. [Accessed 06-08-2024].
28. Carballal, A. *et al*. Automatic multiscale vascular image segmentation algorithm for coronary angiography. *Biomedical Signal Processing and Control* **46**, 1–9, https://doi.org/10.1016/j.bspc.2018.06.007 (2018).
29. Angiographics — figshare.com. https://figshare.com/s/4d24cf3d14bc901a94bf. [Accessed 06-08-2024].
30. Chowdhury, M. E. H. *et al*. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access* **8**, 132665–132676, https://doi.org/10.1109/access.2020.3010287 (2020).
31. Rahman, T. *et al*. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in Biology and Medicine* **132**, 104319, https://doi.org/10.1016/j.compbiomed.2021.104319 (2021).
32. COVID-19 Radiography Database — kaggle.com. https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database. [Accessed 06-08-2024].
33. Tahir, A. M. *et al*. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in Biology and Medicine* **139**, 105002, https://doi.org/10.1016/j.compbiomed.2021.105002 (2021).
34. A M. Tahir *et al*. Covid-qu-ex dataset, https://doi.org/10.34740/KAGGLE/DSV/3122958 (2022).
35. Morozov, S. P. *et al*. Mosmeddata: Chest ct scans with covid-19 related findings dataset, https://doi.org/10.48550/ARXIV.2005.06465 (2020).
36. COVID-19 CT scan lesion segmentation dataset — kaggle.com. https://www.kaggle.com/datasets/maedemaftouni/covid19-ct-scan-lesion-segmentation-dataset. [Accessed 06-08-2024].
37. Garcia-Peraza-Herrera, L. C. *et al*. Image compositing for segmentation of surgical tools without manual annotations. *IEEE Transactions on Medical Imaging* **40**, 1450–1460, https://doi.org/10.1109/tmi.2021.3057884 (2021).
38. Zeeshan Ahmed, Munawar Ahmed, Attiya Baqai & Fahim Aziz Umrani. Intraretinal cystoid fluid, https://doi.org/10.34740/KAGGLE/DS/2277068 (2022).
39. Ahmed, Z. *et al*. Deep learning based automated detection of intraretinal cystoid fluid. *International Journal of Imaging Systems and Technology* **32**, 902–917, https://doi.org/10.1002/ima.22662 (2021).
40. Cervantes-Sanchez, F., Cruz-Aceves, I., Hernandez-Aguirre, A., Hernandez-Gonzalez, M. A. & Solorio-Meza, S. E. Automatic segmentation of coronary arteries in x-ray angiograms using multiscale analysis and artificial neural networks. *Applied Sciences* **9**, 5507, https://doi.org/10.3390/app9245507 (2019).
41. Ivan Cruz Aceves CIMAT — personal.cimat.mx. http://personal.cimat.mx:8181/ivan.cruz/DB_Angiograms.html. [Accessed 06-08-2024].
42. Spahn, C. *et al*. Deepbacs for multi-task bacterial image analysis using open-source deep learning approaches. *Communications Biology* **5**, https://doi.org/10.1038/s42003-022-03634-z (2022).
43. Spahn, C. & Heilemann, M. Deepbacs – escherichia coli bright field segmentation dataset, https://doi.org/10.5281/zenodo.5550935 (2021).
44. Staal, J., Abramoff, M., Niemeijer, M., Viergever, M. & van Ginneken, B. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging* **23**, 501–509, https://doi.org/10.1109/tmi.2004.825627 (2004).
45. DRIVE - Grand Challenge — drive.grand-challenge.org. https://drive.grand-challenge.org/. [Accessed 06-08-2024].
46. Van Valen, D. A. *et al*. Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. *PLOS Computational Biology* **12**, e1005177, https://doi.org/10.1371/journal.pcbi.1005177 (2016).
47. DeepCell Datasets — datasets.deepcell.org. https://datasets.deepcell.org/data. [Accessed 06-08-2024].
48. Lu, Y. *et al*. The jnu-ifm dataset for segmenting pubic symphysis-fetal head. *Data in Brief* **41**, 107904, https://doi.org/10.1016/j.dib.2022.107904 (2022).
49. Jieyun, B. & ZhanHong, O. Pubic symphysis-fetal head segmentation and angle of progression, https://doi.org/10.5281/ZENODO.7851338 (2024).
50. Porwal, P. *et al*. Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. *Data* **3**, 25, https://doi.org/10.3390/data3030025 (2018).
51. Prasanna Porwal, S. P. Indian diabetic retinopathy image dataset (idrid), https://doi.org/10.21227/H25W98 (2018).
52. Codella, N. C. F. *et al*. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, https://doi.org/10.1109/isbi.2018.8363547 (IEEE, 2018).
53. ISIC Challenge — challenge.isic-archive.com. https://challenge.isic-archive.com/data/#2016. [Accessed 07-08-2024].
54. Tschandl, P., Rosendahl, C. & Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* **5**, https://doi.org/10.1038/sdata.2018.161 (2018).
55. Codella, N. *et al*. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), https://doi.org/10.48550/ARXIV.1902.03368 (2019).
56. ISIC Challenge — challenge.isic-archive.com. https://challenge.isic-archive.com/data/#2018 (2018). [Accessed 07-08-2024].
57. Jha, D. *et al*. *Kvasir-SEG: A Segmented Polyp Dataset*, 451–462 (Springer International Publishing, 2019).
58. Simula Datasets - Kvasir SEG — datasets.simula.no. https://datasets.simula.no/kvasir-seg/. [Accessed 06-08-2024].
59. Maqbool, S., Riaz, A., Sajid, H. & Hasan, O. m2caiseg: Semantic segmentation of laparoscopic images using convolutional neural networks, https://doi.org/10.48550/ARXIV.2008.10134 (2020).
60. m2caiSeg — kaggle.com. https://www.kaggle.com/datasets/salmanmaq/m2caiseg. [Accessed 07-08-2024].
61. Verma, R. *et al*. Monusac2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Transactions on Medical Imaging* **40**, 3413–3423, https://doi.org/10.1109/tmi.2021.3085712 (2021).
62. MoNuSAC 2020 - Grand Challenge — monusac-2020.grand-challenge.org. https://monusac-2020.grand-challenge.org/Data/. [Accessed 07-08-2024].
63. Janowczyk, A. & Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics* **7**, 29, https://doi.org/10.4103/2153-3539.186902 (2016).
64. Yang, L. *et al*. Nuset: A deep learning tool for reliably separating and analyzing crowded cells. *PLOS Computational Biology* **16**, e1008193, https://doi.org/10.1371/journal.pcbi.1008193 (2020).
65. Linfeng Y. Nuset training dataset/model weights from (nuset: A deep learning tool for reliably separating and analyzing crowded cells), https://doi.org/10.5281/ZENODO.3996369 (2020).
66. Abdi, A. H., Kasaei, S. & Mehdizadeh, M. Automatic segmentation of mandible in panoramic x-ray. *Journal of Medical Imaging* **2**, 044003, https://doi.org/10.1117/1.jmi.2.4.044003 (2015).
67. Abdi, A. Panoramic dental x-rays with segmented mandibles, https://doi.org/10.17632/HXT48YK462.1 (2017).

68. Ali, S. *et al*. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge, https://doi.org/10.48550/ARXIV.2202.12031 (2022).
69. Ali, S. *et al*. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Medical Image Analysis* **70**, 102002, https://doi.org/10.1016/j.media.2021.102002 (2021).
70. Litjens, G. *et al*. Evaluation of prostate segmentation algorithms for mri: The promise12 challenge. *Medical Image Analysis* **18**, 359–373, https://doi.org/10.1016/j.media.2013.12.002 (2014).
71. Litjens, G. *et al*. Promise12: Data from the miccai grand challenge: Prostate mr image segmentation 2012, https://doi.org/10.5281/ZENODO.8014040 (2023).
72. Jack, N. P., Thomas, W., Laé M. & Reyal F. Segmentation of nuclei in histopathology images by deep regression of the distance map, https://doi.org/10.5281/ZENODO.1175282 (2018).
73. Naylor, P., Laé, M., Reyal, F. & Walter, T. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Transactions on Medical Imaging* **38**, 448–459, https://doi.org/10.1109/tmi.2018.2865709 (2019).
74. Ultrasound Nerve Segmentation — kaggle.com. https://www.kaggle.com/competitions/ultrasound-nerve-segmentation. [Accessed 07-08-2024].
75. Song, Y. *et al*. Ct2us: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. *Ultrasonics* **122**, 106706, https://doi.org/10.1016/j.ultras.2022.106706 (2022).
76. CT2USforKidneySeg — kaggle.com. https://www.kaggle.com/datasets/siatsyx/ct2usforkidneyseg/data. [Accessed 07-08-2024].
77. Skin Cancer Detection | Vision and Image Processing Lab — uwaterloo.ca. https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection. [Accessed 07-08-2024].
78. Zheng, X., Wang, Y., Wang, G. & Liu, J. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron* **107**, 55–71, https://doi.org/10.1016/j.micron.2018.01.010 (2018).
79. Acevedo, A., Alférez, S., Merino, A., Puigví, L. & Rodellar, J. Recognition of peripheral blood cell images using convolutional neural networks. *Computer Methods and Programs in Biomedicine* **180**, 105020, https://doi.org/10.1016/j.cmpb.2019.105020 (2019).
80. Dietler, N. *et al*. A convolutional neural network segments yeast microscopy images with high accuracy. *Nature Communications* **11**, https://doi.org/10.1038/s41467-020-19557-4 (2020).
81. Data and Software — epfl.ch. https://www.epfl.ch/labs/lpbs/data-and-software/. [Accessed 07-08-2024].
82. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *CoRR* **abs/1512.03385** (2015).
83. Kugelman, J. *et al*. A comparison of deep learning u-net architectures for posterior segment oct retinal layer segmentation. *Scientific Reports* **12**, https://doi.org/10.1038/s41598-022-18646-2 (2022).
84. Cinar, N., Ozcan, A. & Kaya, M. A hybrid densenet121-unet model for brain tumor segmentation from mr images. *Biomedical Signal Processing and Control* **76**, 103647, https://doi.org/10.1016/j.bspc.2022.103647 (2022).
85. Aydin, M. & Kuş, Z. Medsegbench: Model weights and predictions, https://doi.org/10.5281/ZENODO.13381081 (2024).

Acknowledgements

We would like to express our gratitude to the authors of the MedMNIST[8], which served as the baseline for our study and for the shared source code we referenced to develop our own code.

Author contributions

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

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.