ÇANKAYA UNIVERSITY
COMPUTER ENGINEERING
DEPARTMENT

CENG 407

Dataset Description & Preprocessing

Dataset Description and Preprocessing

Team Members:

Alperen Berke Çetinkaya

Muhammed Yusuf Özcan

Sezer Ataş

Mete Serpil

Contents

1. Dataset Description

1.1 Use of the BraTS 2020 Dataset in the MedVisVR Project

The BraTS 2020 dataset is an authoritative resource considered the "gold standard" in the field for the automated analysis and segmentation of brain tumors within multimodal Magnetic Resonance Imaging (MRI) scans.

Within the MedVisVR project, this dataset is utilized for the precise detection of glioma-derived pathologies, including High-Grade Gliomas and Low-Grade Gliomas. The primary objective of the project is to transform the high-accuracy segmentation data derived from this dataset into clinically meaningful, 3D interactive models within a Virtual Reality environment.

Purpose of Use and Segmentation Objectives

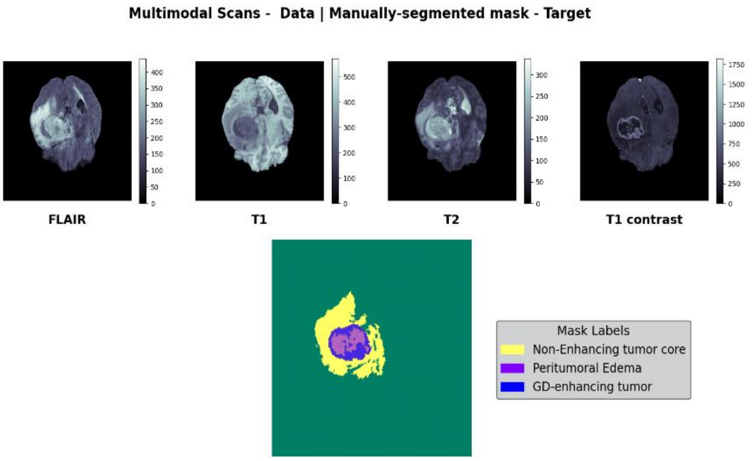
Within the scope of the project, BraTS 2020 data is processed to achieve the automated decomposition of heterogeneous glioma sub-components across multimodal MRI sequences.

The resulting segmentation outputs define the anatomical boundaries, volume, and internal structure of the tumor, forming the foundation for the 3D pathological objects generated on the MedVisVR platform. This process allows clinicians to examine the tumor structure in depth and through multiple layers.

Imaging Data Characteristics

The dataset consists of pre-operative MRI scans collected from 19 different institutions, utilizing various clinical protocols and diverse MRI scanners.

1. **Modalities** Each patient case includes four co-registered MRI sequences:
   - **T1:** Native T1-weighted.
   - **T1ce**: Contrast-enhanced T1-weighted (Gadolinium - T1Gd).
   - **T2:** T2-weighted.
   - **FLAIR:** T2-weighted, Fluid-Attenuated Inversion Recovery.



2. **Format and Pre-processing** : To ensure model stability and data consistency, the raw data has been standardized through the following procedures:

   - **Co-registration:** All modalities have been aligned to the same anatomical space.
   -**Resolution:** Images have been resampled to a 1 mm³ isotropic.

-Skull-stripping: Non-brain tissues and the skull have been removed to focus exclusively on intracranial structures.

-File Format: The original data is provided in NIfTI (.nii.gz) format.

-Optimization: Within the MedVisVR architecture, volume data is converted and utilized in HDF5 (.h5) format to increase memory efficiency during training and visualization processes.

3. Data Volume

   -**Training Set:** 369 cases.

   -**Validation Set:** 125 cases.

MedVisVR Segmentation and Visualization Approach

The standard labels in the BraTS dataset have been restructured within MedVisVR to provide clinicians with the clearest possible imagery in a VR environment:

-Tumor Core: The necrotic/non-enhancing regions and the active enhancing tumor tissue are merged and processed as a single structure. In the VR environment, this is visualized as a solid, holistic object representing the primary mass of the tumor.

-Peritumoral Edema: The edema tissue surrounding the tumor is maintained as a separate layer. During visualization, it is presented with distinct color coding or transparency settings to demonstrate the spread and impact of the tumor on healthy tissue.

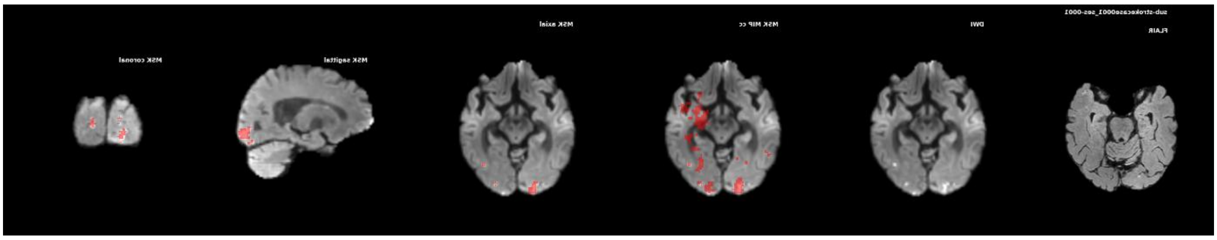
This approach ensures that the relationship between the tumor core and its infiltration into surrounding tissues can be intuitively distinguished during surgical planning or educational simulations.

1.2 Use of the ISLES 2022 Dataset in the MedVisVR Project

The ISLES 2022 (Ischemic Stroke Lesion Segmentation Challenge) dataset is a reliable reference dataset featuring multi-institutional, expert-level annotations focused on lesion segmentation in acute ischemic stroke cases. In the MedVisVR project, this dataset was utilized for the accurate detection and volumetric analysis of ischemic lesions (infarct core), as well as their evaluation in 3D within the context of brain anatomy in a VR environment.

Dataset Content and Imaging Modalities

The ISLES 2022 dataset includes three fundamental MR sequences for each case, accompanied by expert-generated ischemic lesion annotations: FLAIR, Diffusion-Weighted Imaging (DWI), and corresponding ADC maps. These modalities provide critical clinical information, particularly in defining the infarct core (dead tissue) and clarifying lesion boundaries during the acute stroke process.



Data Format and Pre-processing

All images and annotations were provided in NIfTI format in compliance with the **BIDS (Brain Imaging Data Structure)** standard, which is widely used in brain imaging studies. Images were shared in the patients' native space, and no additional spatial alignment was applied. To protect patient privacy, skull-stripping was performed on all data. For applicable cases, DICOM metadata was provided in JSON format to ensure the traceability of imaging parameters.

Data Source and Imaging Centers

The dataset was collected retrospectively from three different stroke centers using MR systems with varying magnetic field strengths. This diversity allowed for the evaluation of the robustness of the methods developed within MedVisVR against different devices and clinical scenarios.

Annotation Structure

Within the scope of ISLES 2022, all cases were labeled under a single class: ischemic lesion (infarct core). These annotations were created by expert clinicians, providing a reliable foundation for segmentation accuracy.

Implementation Approach in MedVisVR

In the MedVisVR project, ISLES 2022 data was integrated for the precise segmentation of ischemic lesions using diffusion-based MR images. The resulting segmentation results were converted into 3D models to reveal the spatial location and volume of the lesion within the brain anatomy. In the VR environment, ischemic lesions are represented with a blue color code to facilitate visual differentiation.

Through this method, the aim is to enable clinicians to evaluate the spread, boundaries, and relationship of the stroke lesion with brain tissue in a more intuitive and holistic way.

Data Volume

The ISLES 2022 dataset used in the project consists of approximately 250 expert-annotated acute ischemic stroke cases. This data served as the primary input for training the lesion segmentation model and developing the MedVisVR visualization components.

1.3 Use of the TotalSegmentator Dataset in the MedVisVR Project

General Characteristics of the Dataset

TotalSegmentator is a comprehensive anatomical segmentation dataset created from large-scale, whole-body computed tomography (CT) scans. Within MedVisVR, it provided an infrastructure that strengthens the relationship between neurological pathologies and surrounding tissues, specifically by taking the head and upper body anatomy as a reference.

The dataset covers versions v1 and v2 published via the Zenodo platform. The first version includes 1228 CT scans, while version v2 has expanded data diversity and annotation scope. All data is provided in NIfTI format, making it compatible with MedVisVR's image processing and visualization components.

Anatomical Labels and Scope

The TotalSegmentator dataset includes detailed labels covering 104 different anatomical structures, including bone structures, organs, vessels, and soft tissues. This rich annotation structure enabled the alignment of pathological segmentations with anatomical references in MedVisVR, presenting them in a more clinically meaningful context.

Implementation Approach in MedVisVR

In the MedVisVR project, TotalSegmentator data was used independently of pathology-focused segmentations (e.g., tumors or ischemic lesions) to create an anatomical framework. Thanks to this framework, the pathological structures shown in the VR environment were placed in the correct position and scale within a realistic head anatomy.

This approach allows users to intuitively evaluate not only the pathology itself but also its relationship with surrounding anatomical structures in 3D.

1.4 Use of the CHAOS – Combined (CT-MR) Healthy Abdominal Organ Dataset in the MedVisVR Project

The CHAOS – Combined (CT-MR) Healthy Abdominal Organ dataset is a comprehensive reference dataset prepared for abdominal organ segmentation, containing both CT and MR images. The dataset consists of images obtained from healthy individuals, allowing for the detailed and accurate modeling of abdominal anatomy. In this regard, it provides a reliable basis for determining organ boundaries, volumetric analysis, and method evaluation in clinical applications.

In the MedVisVR project, the CHAOS dataset was not used directly for pathological segmentation; instead, it was evaluated as a supporting resource for creating anatomical context and visually defining organ boundaries. Specifically, referencing the torso anatomy made it possible to present pathological structures in a more meaningful and realistic context within the VR environment.

Dataset Structure and Content

The CHAOS dataset includes abdominal CT and MR images obtained from different individuals. Both imaging modalities focus on normal organ structures, i.e., those containing no pathological findings. This allows for the clear representation of the anatomical boundaries of primary abdominal organs such as the liver, kidneys, and spleen.

Format and Pre-processing

While the dataset was originally provided in clinical imaging formats, the images were converted to NIfTI (.nii.gz) format within the MedVisVR project to align with standard image processing and visualization workflows. The data structure was organized with a training and testing split; the training data includes both the images and their corresponding manual segmentation masks.

Organ Labels

Labeling in the CHAOS dataset covers different organs depending on the modality. While CT images primarily feature liver segmentation, MR images contain segmentations of multiple abdominal organs, including the liver as well as the right and left kidneys and the spleen. This structure allows for the modeling of anatomical relationships between different organs.

Implementation Approach in MedVisVR

In the MedVisVR project, healthy organ segmentations obtained from the CHAOS dataset were used to create 3D anatomical models. These models were integrated into the VR environment to show the spatial relationship of pathological segmentations with surrounding organs. Thus, pathological structures were positioned within a realistic anatomical environment rather than as abstract objects.

This approach allows users to evaluate not only the pathology itself but also its relationship with surrounding organs in a holistic manner. In this way, the CHAOS dataset has contributed to the project as an important component that strengthens the anatomical consistency of MedVisVR and enriches the visualization experience.

1.5 CT-ORG (CT Volumes with Multiple Organ Segmentations)

The CT-ORG dataset, used to ensure that the MedVisVR project masters the entirety of human anatomy rather than just specific pathologies (tumors, hemorrhages, etc.), is the fundamental building block that enhances the model's "scene perception." Compiled by The Cancer Imaging Archive (TCIA), this collection is used to enable our AI model to distinguish multiple organ structures in different body regions simultaneously and with high accuracy (Multi-Organ Segmentation).

1. Project Scope and Strategic Importance

In real-world scenarios, medical images are not standardized; patient posture, weight, or organ placement can vary. CT-ORG plays a critical role in training our model against these variations.

- The primary function of this dataset in the project is:
- **Contextual Learning:** It enables the model to not only find the "liver" but to understand the liver's proximity to the lungs and bone structures. Consequently, organs are visualized in the VR environment not as disjointed parts, but as an anatomical whole (like a completed jigsaw puzzle).

2. Data Source and Diversity (Heterogeneity)

Instead of data from a single clinical study, CT-ORG consists of 140 CT series compiled from various sources (e.g., the Liver Tumor Segmentation Challenge - LiTS, etc.).

This "mixed" structure is a major advantage for our project:

- **Generalizability:** Since the data comes from different CT devices and centers, our model does not "overfit" to the image quality or noise type of a single hospital. Instead, it gains a robust structure capable of recognizing organ boundaries even under varying imaging conditions.

3. Targeted Anatomical Structures

The dataset includes segmentation masks for the following radiologically critical organs covering a large portion of the body (from the chest to the pelvis):

1. **Lungs:** To define the boundaries of the thoracic region and create a respiratory system reference.
2. **Liver:** As the most dominant organ of the abdomen, serving as the central point of abdominal navigation.
3. **Kidneys:** Critical references for modeling the retroperitoneal (behind the abdominal lining) region.
4. **Bladder:** Defines the lower boundary of the pelvic region.
5. **Bones:** Forming the skeletal frame in which all these soft tissues are placed, providing the basic reference for the user's orientation in the VR environment.
6. **Brain***:* Available in some series, used to ensure integrity in head region scans.

4. Data Processing and Standardization

In their raw form, CT-ORG data have different slice thicknesses and resolutions. Before being fed into the MedVisVR "Anatomical Context Engine," they underwent the following processes:

- **Hounsfield Unit (HU) Windowing: s**Density ranges where different organs (soft tissue vs. bone) are most clearly visible were normalized. This allows the model to focus on tissue characteristics independently of the image's brightness/contrast settings.
- **Format Uniformity:** All volumes and masks were converted to the project standard, NIfTI (.nii.gz) format.
- **Quality Control:** Since the masks in the dataset were produced with automated methods and corrected by experts (hybrid annotation), they underwent a final check for anatomical consistency before training.

1.6 VerSe'20

The **VerSe'20** dataset is one of the primary reference datasets ensuring that pathological structures are placed within a realistic anatomical framework in the MedVisVR system. Specifically, in terms of modeling the head and upper spine anatomy, this dataset, which focuses on spinal segmentation and vertebrae labeling tasks was used to enhance the accuracy of the anatomical context.

Dataset Purpose and Scope

VerSe'20 is a large-scale dataset created from CT images with the goals of:

- Automated segmentation of the vertebrae forming the spine.
- Accurate labeling of the anatomical level of each vertebra.

In the context of MedVisVR, this dataset enables pathological segmentations to be presented within a clinically meaningful skeletal and head anatomy, rather than as isolated objects in space.

Imaging Modality and Format

- **Modality:**
  Computed Tomography (CT)
- **Format:**
  All images and segmentation masks are provided in **NIfTI (.nii / .nii.gz)** format.

This format is fully compatible with the MedVisVR pre-processing pipeline, allowing for the direct transfer of 3D volumes into the VR environment.

Annotations and Labeling Structure

For each case, the VerSe'20 dataset includes the following components:

- Voxel-based vertebrae segmentation masks, allowing for the 3D decomposition of each vertebra in the spine.

- Anatomical level labels defining the positions of the vertebrae in the cervical, thoracic, and lumbar regions.

Through this structure, spinal anatomy can be modeled semantically as well as geometrically.

This anatomical structure is integrated through a fusion process with:

Ischemic lesions derived from ISLES 2022.

Consequently, the spatial relationship between these pathologies and the cranial and spinal anatomy is presented with high fidelity within the VR environment. This approach ensures that the pathological findings are not viewed in isolation but are contextualized within the patient's actual anatomical framework.

Systemic Contribution

The primary contributions of VerSe’20 to the MedVisVR architecture are as follows:

- Preventing the abstract visualization of pathological structures without anatomical references.
- Reducing errors in scale, position, and orientation within the VR environment.
- Enabling clinical users to evaluate the pathology in relation to the skeletal and head anatomy.

In this way, MedVisVR evolves from a system focused solely on segmentation accuracy into a context-aware visualization platform for clinical decision support.

2. Dataset Tables

2.1 Primary Research Datasets

Dataset	Imaging Modality	Target Structures (Labels)	Primary Role in MedVisVR	Data Format
BraTS 2020	**Multimodal MRI** (T1, T1ce, T2, FLAIR)	**Glioma Tumor Sub-regions:** • Tumor Core (NCR/NET + ET) •Peritumoral Edema	**Pathological Visualization:** Detection and segmentation of brain tumors and their presentation as 3D pathological objects within the VR environment.	NIfTI (Converted to HDF5)
ISLES 2022	**MRI** (DWI, ADC, FLAIR)	**Ischemic Lesion:** • Infarct Core (Dead Tissue)	**Pathological Visualization:** Delineation of lesion boundaries in acute stroke cases and volumetric representation using a blue color code.	NIfTI (BIDS Standard)
TotalSegmentator (v1 & v2)	**CT**	**Anatomical Structures:** • Skull • Vertebrae • Skin	**Anatomical Context:** Creating a reference skeleton to position pathologies within a realistic head/body structure rather than in empty space.	NIfTI
CHAOS	**CT & MRI**	**Abdominal Organs:** • Liver • Kidneys • Spleen	**Anatomical Context:** Establishing spatial relationships between pathologies and surrounding tissues by referencing healthy organ boundaries.	NIfTI
CT-ORG	**CT**	**Multi-Organs:** • Lungs, Liver, Bone Structure, Bladder, Brain	**Scene Perception & Generalizability:** Enabling the model to learn "anatomical integrity" (scene perception) against variations in patient posture and device differences.	NIfTI

Dataset	Imaging Modality	Target Structures (Labels)	Primary Role in MedVisVR	Data Format
VerSe'20	**CT**	**Spine:** • Vertebra Segmentation •Anatomical Level Labels	**Anatomical Context:** Ensuring proper orientation, particularly in the head and neck region, and modeling the spinal anatomy.	NIfTI

2.2 External Validation & Reference Datasets

Dataset	Imaging Modality	Target Structures (Labels)	Primary Role / Intended Use	Data Format
MSD (Medical Segmentation Decathlon)	**Multimodal** **(CT, MRI - T1, T2, FLAIR, ADC)**	**10 Different Organs & Pathologies:** • Brain (Glioma) • Heart (Atrium) • Liver & Tumor •Hippocampus • Prostate • Lung • Pancreas • Hepatic Vessel • Spleen • Colon	**Generalizability Testing:** Developing and comparing "general-purpose" segmentation models capable of working across different anatomical structures and modalities.	**NIfTI** (With standardized directory structure and JSON metadata)
TCIA – Glioma, Head-Neck, Lung	**Multimodal** (MRI, CT, PET, Pathology)	**Cancer-Focused Structures:** • Glioma: Tumor sub-regions (Necrosis, Edema) • Head-Neck: Tumor volume and organs at risk • Lung: Nodules and masses	**Clinical Research & Radiomics:** Treatment response assessment in real clinical scenarios, radiotherapy planning, and genomic data matching.	**DICOM** (Raw clinical image format, requires processing)

Dataset	Imaging Modality	Target Structures (Labels)	Primary Role / Intended Use	Data Format
Brain MRI Dataset (Multiple Sclerosis)	**MRI** (T1, T2, FLAIR)	**Pathological Lesions:** • MS (Multiple Sclerosis) Plaques • White Matter Lesions	**Disease Monitoring & Training:** Model training for detection and volumetric analysis of MS lesions, and correlation with patient disability status (EDSS).	**NIfTI** (Usually provided in processed or raw format)

3. Preprocessing And Standardization Pipeline

3.1 Executive Summary

This report describes the data preprocessing pipeline designed for the BraTS 2020 dataset, which contains multi-modal MRI scans. The goal of this pipeline is to prepare the data for training 3D Convolutional Neural Networks specifically 3D U-Net architecture. The workflow is optimized to handle hardware limitations, reduce data heterogeneity, and mitigate class imbalance issues inherent in tumor segmentation tasks.

3.2. Technical Specifications (Input/Output Overview)

Parameter	Raw Data (Input)	Processed Data (Output)
Dimensions (Tensor Shape)	(240, 240, 155)	(128, 128, 128) (Patch-Based)
Channel Structure	4 Separate Files (T1, T1ce, T2, FLAIR)	(4, 128, 128, 128) (Stacked Tensor)
Labels	0, 1, 2, 4 (Original Classes)	3-Channel Multi-Label (WT, TC, ET)
Voxel Resolution	Variable (may be anisotropic)	1.0 × 1.0 × 1.0 mm (Isotropic)
Orientation	Inconsistent	RAS (Right–Anterior–Superior)
Pixel Intensities	Raw MRI Intensity (0–3000+)	Normalized (Mean = 0, Std = 1)

3.3 Processing Pipeline

The preprocessing architecture is implemented using the MONAI library and follows this sequential workflow:

1. **Data Loading**
   Loading and stacking the four MRI modalities.
2. **Label Mapping**
   Converting tumor labels into clinical subregions (WT / TC / ET).
3. **Spatial Transforms**
   - Orientationd(RAS): Standardizes image orientation.
   - Spacingd(1mm): Resamples all scans to 1 mm isotropic spacing.
4. **Intensity Normalization**
   Z-score normalization applied only to brain tissue.
5. **Class-Balanced Patch Extraction**
   Using RandCropByPosNegLabeld to ensure patches with tumor regions are preferentially sampled.
6. **Data Augmentation**
   Random flips and rotations to increase data diversity.
7. **Export**
   Writing the processed tensors back to .nii files.

3.4 Methodology Details

3.4.1. Multi-Modality and Semantic Transformation

A multi-modal learning strategy is used to help the model better capture tumor boundaries. Original labels are reorganized into three clinically relevant regions:

- **Whole Tumor (WT)**
- **Tumor Core (TC)**
- **Enhancing Tumor (ET)**

This multi-label representation is known to improve segmentation performance, especially Dice scores.

3.4.2. Computational Strategy: Label-Guided Sampling

The raw MRI volumes (240 × 240 × 155) are large and memory-intensive. Additionally, tumor regions occupy only a small fraction of the volume, making naïve random cropping ineffective most patches contain only background brain tissue.

To address this, the pipeline uses **RandCropByPosNegLabeld**, which ensures:

- Training volumes are cropped into 128 × 128 × 128 patches.
- Each batch contains patches in a **2:1 ratio of tumor-containing vs. healthy tissue**.
- Tumor regions are consistently included during training, significantly reducing class imbalance.

3.5 Error Handling and Quality Control

To ensure robustness, the following safeguards are implemented:

- Missing File Detection: The pipeline verifies file integrity before processing.
- Channel Consistency: EnsureChannelFirstd enforces PyTorch's (C, H, W, D) channel ordering.
- Determinism: set_determinism (seed=0) ensures reproducible augmentation and transformations.

3.6 Conclusion

The revised preprocessing pipeline produces a dataset that is not only technically standardized but also optimized for high-performance tumor segmentation. Through normalization, orientation correction, label refinement, and class-balanced sampling, the pipeline effectively prepares MRI data for training 3D U-Net models and enhances overall Dice score performance.

4. Ethical Considerations and Data Privacy

4.1. Use of Open-Access Research Repositories

The MedVisVR project is built upon established, open-source medical datasets (BraTS 2020, ISLES 2022, TotalSegmentator, CT-ORG, VerSe'20, and CHAOS). These datasets are hosted by globally recognized academic and clinical repositories, such as The Cancer Imaging Archive (TCIA) and Zenodo. By utilizing these pre-existing, publicly available resources, the project ensures that:

Ethical Oversight: All data was originally collected under the approval of the respective Institutional Review Boards (IRBs) of the contributing hospitals.

Conflict Mitigation: Using open-access data eliminates ethical conflicts regarding primary data collection and patient consent, as the data has already been cleared for international research and development purposes.

4.2. Adherence to Licensing and Usage Policies

The project strictly complies with the specific licensing terms of each dataset (e.g., Creative Commons Attribution 4.0 International - CC BY 4.0). All data is used solely for the stated research, development, and educational goals of MedVisVR. In accordance with competition rules (such as those of the BraTS and ISLES challenges), the project ensures that any comparative results or findings are reported transparently and ethically.

4.3 Scientific Integrity and Secondary Data Use

As a secondary data user, the MedVisVR team maintains scientific integrity by referencing all original authors and institutions. The reliance on these "gold standard" benchmarks ensures that the project's results are reproducible and comparable with other state-of-the-art methods in the medical imaging community.

5. References

1.https://www.kaggle.com/datasets/awsaf49/brats20-dataset-training-validation

2.https://www.kaggle.com/datasets/orvile/isles-2022-brain-stoke-dataset

3.https://zenodo.org/records/10047292

4.https://zenodo.org/records/14710732

5. https://www.kaggle.com/datasets/omarxadel/chaos-combined-ct-mr-healthy-abdominal-organ

6.https://www.cancerimagingarchive.net/collection/ct-org/

7.https://s3.bonescreen.de/public/VerSe-complete/dataset-verse20test.zip

8. http://medicaldecathlon.com/

9. https://www.cancerimagingarchive.net/brain-and-head-neck-imaging-still-available-on-tcia/

10. https://www.kaggle.com/datasets/trainingdatapro/multiple-sclerosis-dataset