Automated Detection of Craniomaxillofacial Fractures From 3D CT Images Using Ensemble Deep Learning-Based Segmentation Models

Muhammad Febrian Rachmadi, Prasetyanugraheni Kreshanti, Margareth Ingrid Anggraeni, Vika Tania, Reyhan Eddy Yunus, and Henrik Skibbe

***Abstract*— Craniomaxillofacial (CMF) fractures, often resulting from traffic accidents, falls, and head traumas, necessitate prompt diagnosis and analysis with CT images. Our study leverages a segmentation model named 3D Swin UNETR to develop an automated detection system for these fractures. The key finding of this study is the significant improvement in the quality of CMF fracture detection achieved by incorporating an additional input channel containing labels of skull regions, using an additional loss function named Proximity loss, and performing an ensemble inference approach using different models trained by different settings. Clinical evaluations were manually performed by experts where the best-performing model achieved the positive predictive value (PPV) of 82.49%, true positive rate (TPR) of 96.03%, false detection rate (FDR) of 17.51%, false negative rate (FNR) of 3.97%, and F1-score (F1) of 88.23%.**

***Index Terms*— Craniomaxillofacial injuries, detection, segmentation, deep learning, CT images.**

I. Introduction

CLINICALLY known as craniomaxillofacial (CMF) fractures, facial fractures are commonly caused by traffic accidents, bicycle accidents, industrial accidents, assaults, domestic violence and sports injuries. Airway obstruction is one of the serious complications, where this is exacerbated by the risk of aspiration and vomiting [1]. Additionally, CMF fractures resulting from a significant impact to the head can have a negative effect on the patient's prognosis because they can cause intracranial hematomas and infections [2]. Given the potential for early yet advanced complications leading to suboptimal outcomes, prompt and precise initial diagnosis is crucial [3], [4].

The computed tomography (CT) scan of the patient's head is considered the gold standard for diagnosing CMF fractures [5]. However, interpreting and diagnosing CMF fractures via CT scan is challenging due to the complex anatomical structure of CMF. For example, some CMF structures interdigitate with each other causing CMF fractures to disrupt adjacent structures in complex ways [6], often resulting in more complex CMF fractures. Because of that, CMF fractures can appear in several regions of the skull simultaneously (e.g., the mandible, maxilla, skull base) which might complicate the diagnosis. Therefore, diagnosing CMF fractures via CT scan by human experts requires a lot of experience, precision, effort, and time [3], [7] and is often prone to unintended omissions and misjudgments [7]–[9].

Based on the challenges discussed above, this study developed an automated detection of CMF fractures from whole 3D head CT scan images using ensembles of deep segmentation neural networks (i.e., deep learning). Deep learning itself has been increasingly utilized in various medical fields over the past few years, exhibiting satisfactory performance in medical image positioning, segmentation, and diagnosis [10]. **Our contributions in this study are** (1) we presented the first study on the automatic detection of CMF fractures from full 3D volume images of CT scans using deep segmentation neural networks, (2) we proposed a unique approach of incorporating labels of skull regions, using a compound loss function for training, and using ensembles of segmentation models to improve the quality of CMF fracture detection, and (3) we performed not only quantitative assessments but also clinical assessments to assess the clinical applicability of our proposed approaches.

The rest of this paper is organized as follows. Section II discusses related previous works on automated detection of fractures on the head. Section III describes the dataset used in this study. Section IV explains all different settings used in our experiments. Section V describes and discusses the results. Lastly, Section VI concludes this study.

II. Related Works

Before the deep learning, only a few previous studies conducted research on the automated detection of skull fractures.

Submitted on... "This work was supported by the program for Brain Mapping by Integrated Neurotechnologies for Disease Studies (Brain/MINDS) from the Japan Agency for Medical Research and Development AMED (JP15dm0207001). Library access provided by Universitas Indonesia is also gratefully acknowledged."

M.F. Rachmadi and H. Skibbe are with the Brain Image Analysis Unit, RIKEN Center for Brain Science, Wako-shi, Japan (e-mail: febrian.rachmadi@riken.jp).

M.F. Rachmadi is also with the Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia (e-mail: febrianrachmadi@cs.ui.ac.id).

P. Kreshanti, M.I. Anggraeni, and V. Tania are medical doctors with the Cleft and Craniofacial Center Dr. Cipto Mangunkusumo Hospital, Division of Plastic Reconstructive and Aesthetic Surgery, Department of Surgery, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia.

R.E. Yunus is a medical doctor with the Department of Radiology, Ciptomangunkusumo Hospital, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia.

M.F. Rachmadi, P. Kreshanti, and R.E. Yunus are also with the Medical Technology IMERI, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia

These studies mostly used computer vision techniques, such as the black-hat transform, to detect fractures on the skull from head CT scans [11], [12]. However, the computer vision techniques were unreliable due to the variety and complexity of the fractures that appear on the skull. In contrast, some previous have proposed several deep learning models to automatically detect skull fractures, which can be categorized into classification, detection, and segmentation approaches.

The majority of previous studies using deep learning proposed classification models to classify whether there are fractures in head CT images with no localization of the fractures. The input image can be a 2D slice of a head CT scan [13], a 2D image/patch of a specific area of the skull (e.g., mandible [14]), or a 3D image/patch of a specific area of the skull (e.g., nasal [15]). One study developed a classification method to classify whether any maxillofacial fractures appear on a 2D slice of a head CT scan or not [16].

On the other hand, other previous studies proposed object detection models by using variations of convolutional neural networks (CNN) [17]–[19], such as Skull R-CNN [18] and Fracture R-CNN [19] models, for skull fractures detection in a 2D slice of a head CT scan. One previous study specifically compared the effectiveness of object detection and semantic segmentation models for detecting fractures in the cranial vaults of the skull, where the semantic segmentation model achieved better detection results [20]. This finding is interesting because fracture segmentation is more difficult than fracture detection. After all, fracture segmentation not only looks for the location but also the shape and size of the fracture [6]. Another study performed a clinical study of an automated diagnosis system for skull fracture detection using a combination of conventional and deep learning algorithms [21].

III. Dataset

A. Data and Subjects

For this study, we collected a retrospective dataset consisting of 20 CT images from 20 patients with CMF fractures admitted to the Dr. Cipto Mangunkusumo Hospital (RSCM), Jakarta, Indonesia in 2020. RSCM is an educational and research hospital for the Faculty of Medicine, Universitas Indonesia. The retrospective data used in this study were approved to be used for research by the Health Research Ethics Committee of the Faculty of Medicine Universitas Indonesia (No. KET-1842/UN2.F1/ETIK/PPM.00.02/2023). The dataset consists of CT images captured by Siemens SOMATOM Definition Flash 128 slice dual source, Philips Ingenuity 64 slice, and Siemens SOMATOM Definition AS 64 slice, which produced CT images with different sizes and voxel's spacings. Characteristic of the subjects is presented in Table I. In this dataset, each subject has CMF fractures with various shapes, sizes, and locations. Therefore, subjects who do not have CMF fractures in a specific region of the skull are beneficial to measure the probability of false detections in that region of the skull.

TABLE I: Baseline characteristic of the subjects.

Characteristic	n (%)
Total patients	20
Male	19 (95)
Female	1 (5)
Age (in years)	
Mean	29.1
Range	6-60
Fracture location	
Frontal cranial vault	2 (10)
Upper central midface	11 (55)
Intermediate central midface	2 (10)
Lower central midface	3 (15)
Maxillary body	12 (60)
Palate	1 (5)
Zygomatic arch	6 (30)
Symphyseal and parasymphyseal mandible	2 (10)
Body mandible	9 (45)
Ramus mandible	1 (5)
Condylar processus mandible	6 (30)

B. Ground Truth Labels

1) Labels of the CMF Fracture: The biggest challenge of this study involved manually creating labels for CMF fractures for all subjects. This labeling was carried out by experts, who delineated both fractures and parts of the skeleton that should have been connected (see yellow boxes in Fig. 1). Note that CMF fractures, unlike other abnormal features in radiology like brain lesions, do not have boundaries that can be easily delineated by experts. Therefore, manual segmentation labels in CMF fractures may extend beyond the skeleton. In addition, as suggested by previous studies [7]–[9], there might be some unintended omissions in the creation of manual labels for CMF fractures by the experts.

2) Labels of the Skull and Skull Regions: Our previous study suggested that global spatial information improved the quality of segmentation tasks in biomedical images when CNN models are used [22]. Therefore, in this study, we proposed to incorporate global spatial information in the form of either labels of the skull or labels of the skull's regions. We generated skull labels for each subject by applying CT image intensity thresholding. Voxels were classified as skull components if their intensities were $\geq$ 600, and then the non-head skeletons that appear in the head CT image (e.g., vertebrae) were manually erased. On the other hand, a trained expert manually labeled the skull into 7 regions, which are mandible, maxilla, zygoma, frontal cranial vault, spheno & temporo lateral, occipital cranial vault, and parietal cranial vault. Examples of 3D visualization of skull and skull region labels are shown in Fig. 2. Note that the labeling processes for CMF fractures, skull, and skull regions were conducted separately.

IV. Experimental Settings

A. Deep Segmentation Model

In this study, we used the 3D Swin UNETR segmentation model [23] from the MONAI library [27] with its default hyperparameter values for our experiments. We chose this model because it was the state-of-the-art deep learning model for segmentation tasks in medical image analysis at the time this study was conducted. Our preliminary experiments have also shown that Swin UNETR outperformed the basic yet

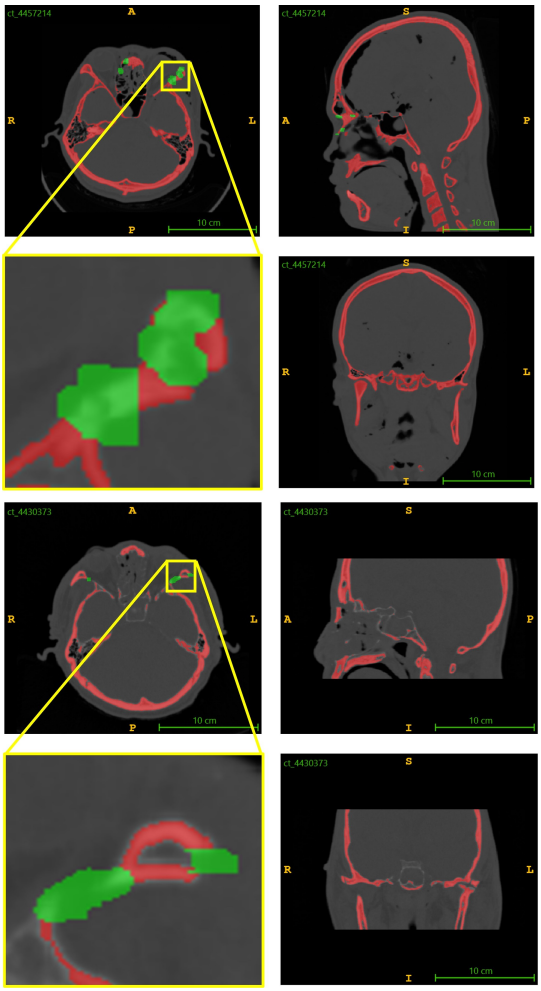

Fig. 1: Visualization of data and labels available for this study, which are CT images and associated labels of the skull (red) and CMF fractures (green). Note that labels for CMF fractures cover not only the gaps between skulls but also some parts of the skull that should have been connected in normal condition (see yellow boxes).

commonly used U-Net model [28]. The visualization of Swin UNETR architecture and how we used the Swin UNETR are illustrated in Fig. 2. Note that we chose to perform the semantic segmentation task over the object detection task because we aimed to not only detect CMF fractures but also segment the lines of CMF fractures and a previous study showed that the semantic segmentation approach achieved better results on the skull fracture detection [20]. In this study, a voxel was predicted to be part of a CMF fracture if the probability value produced by the 3D Swin UNETR segmentation model ≥ 0.5.

B. Training, Validation, and Testing

In this study, we performed a 5-fold nested cross-validation where 20 volumes (from 20 subjects) were randomly divided into 5 groups. In each fold, 3 groups were used for training, 1 group was used for validation (i.e., used for early stopping to avoid overfitting), and 1 group was used for testing. All groups were rotated evenly so that each group was used at least once in the validation and testing sessions. In each fold, a 3D Swin UNETR model was trained for 500 epochs by using the Adam optimizer with a learning rate of 0.0004 and a weight decay of 0.00001. Before the training, we pre-processed each 3D CT image, where its intensities (i.e., Hounsfield units) were scaled from $[-1000, 1000]$ (i.e., to cover both air and various bone densities) to $[0, 1]$. Furthermore, the voxel's dimension/spacing for all subjects was normalized to 1 mm. Data augmentation in the form of random left-and-right flipping (with a probability of 0.75) and random intensity scaling and shifting (with both factor and offset values of 0.5 with a probability of 0.5) were also performed in the training. We conducted our experiments using NVIDIA's a100 GPUs with CUDA version 11.7, Pytorch version 1.13.0, and MONAI version 1.1.0.

C. Input Image Settings

1) Patch-based vs. Full-image Training/Inference: Previous studies discussed in Section II have tested both patch-based and full-image training/inference approaches for skull fracture detection and segmentation with varying degrees of success. However, no previous studies have tried to perform a full 3D volume head CT scan. Therefore, we performed two settings of training/inference: (1) 3D grid patches of head CT scans with a size of $128 \times 128 \times 128$ (i.e., coded as P) and (2) full 3D images of head CT scans with a normalized size of $160 \times 160 \times 160$ (i.e., coded as F). In both settings, weight updates in the training process were performed using mini-batches with a size of 1 patch/image.

2) Number of Input Image Channels: Our previous study indicated that spatial information improved the performance of segmentation tasks in biomedical images [22]. Therefore, we proposed the use of an additional input channel containing either a binary label of the skull (i.e., S) or labels of the skull's regions (i.e., SR) to represent spatial information of the head/skull in a head CT image. Therefore, we performed three different experiments with different input channels for this study, in which we only used CT images (i.e., CT), a combination of CT and S images (i.e., CT+S), and a combination of CT and SR images (i.e., CT+SR).

D. Loss Functions

1) Generalized Dice and Focal Loss: Generalized Dice and Focal (GDF) loss is a compound loss formed of Generalized Dice loss [24] and Focal loss [25]. It was used in our experiments as the default/baseline loss function because it emerged as the best loss function in the preliminary experiments compared to other segmentation loss functions (i.e., Dice loss, Generalized Dice loss [29], and Focal loss). In this study, we used an implementation of the GDF loss function taken from the MONAI library [27] with its default parameters.

2) Instance Proximity Loss: Instance Proximity loss (i.e., Proximity loss) is a novel instance-level loss function proposed in our recent study crafted to refine the detection quality of deep segmentation networks by pulling predicted segmentation instances towards the ground truth instances [26]. Unlike other segmentation losses, the Proximity loss not only calculates the

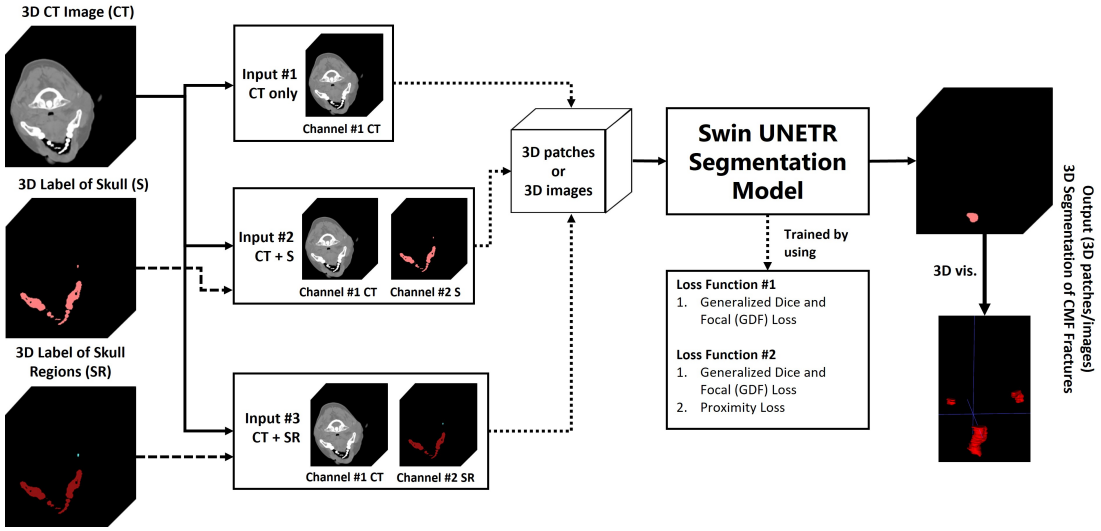

Fig. 2: Illustration of the proposed approach in this study. We tested three options of input to the deep segmentation networks of Swin UNETR [23], which are CT image only, the combination of CT image and binary label of the skull (CT + S), and the combination of CT image and labels of skull's regions (CT + SR). Each input option could be fed into the Swin UNETR either as a 3D patch or a 3D fully sized image of a CT scan. We also tested two options of loss function, which were a compound of the Generalized Dice [24] and Focal [25] (GDF) loss functions and a compound of GDF and Instance Proximity [26] loss functions.

fitness of the intersection of union (IoU) between the predicted segmentation and ground truth masks but also calculates the distance between the closest pairs of each predicted and ground truth instance, which forces a deep segmentation model to produce not only good segmentation but also good instance-level detection.

To improve the detection quality of the segmentation results, the Proximity loss utilizes an object detection loss named Distance-IoU (DIoU) loss [30] which is formalized as

$$\mathcal{L}_{\text{DIoU}}\left(\mathcal{I}_{\hat{y}_c,m},\mathcal{I}_{y_c,n}\right) = 1 - \text{IoU}(\mathcal{I}_{\hat{y}_c,m},\mathcal{I}_{y_c,n}) + \frac{\rho^2\left(\mathfrak{C}_{\mathcal{I}_{\hat{y}_c,m}},\mathfrak{C}_{\mathcal{I}_{y_c,n}}\right)}{\tau^2} \quad (1)$$

where $\mathfrak{C}_{\mathcal{I}_{\hat{y}_c,m}}$ and $\mathfrak{C}_{\mathcal{I}_{y_c,n}}$ denote the centers of the bounding boxes of instances $\mathcal{I}_{\hat{y}_c,m}$ and $\mathcal{I}_{y_c,n}$, $\rho(\cdot)$ is the Euclidean distance function, and τ is the diagonal length of the smallest enclosing box covering the two boxes. If applied to all predicted and ground truth instances, the DIoU loss produces an $N \times M$ distance matrix of $\mathcal{L}_{\text{DIoU}}^{(\mathcal{I}_{\hat{y}_c},\mathcal{I}_{y_c})}$, which is used by the Proximity loss in (3) and (4). This matrix describes the closeness of all N ground truth instances and all M predicted instances with value 0 indicating the maximal closeness.

The Proximity loss itself, which is formalized in (2), can be optimized by minimizing the mean square error (MSE) between two values of $\mathcal{L}_{\text{DIoU}}^{(\text{P})}$ and $\mathcal{L}_{\text{DIoU}}^{(\text{GT})}$. The $\mathcal{L}_{\text{DIoU}}^{(\text{P})}$, formalized in (3), represents the summation of all instance-wise segmentation loss values for each predicted segmentation (P) instance (i.e., $\mathcal{L}_{GDF}^{(\text{P})}$) weighted by a distance value to the closest ground truth (GT) instance calculated by using the DIoU loss function (i.e., the right term of (3)). Similarly, the $\mathcal{L}_{\text{DIoU}}^{(\text{GT})}$, formalized in (4), represents the summation of all instance-wise segmentation loss values for each GT instance (i.e., $\mathcal{L}_{GDF}^{(\text{GT})}$) weighted by a distance value to the closest P segmentation instance calculated by using the DIoU loss (i.e., the right term of (4)). Symbols m and n represent indices for the predicted and ground truth instances. Figure 3 shows the illustration of Proximity loss's calculation on toy images.

$$\mathcal{L}_{\text{proximity}} = \text{MSE}\left(\mathcal{L}_{\text{DIoU}}^{(\text{P})}, \mathcal{L}_{\text{DIoU}}^{(\text{GT})}\right) \quad (2)$$

$$\mathcal{L}_{\text{DIoU}}^{(\text{P})} = \sum_{m}^{M} \mathcal{L}_{GDF}^{(\text{P})}(m) \cdot \min_{n \in N}\left(\mathcal{L}_{\text{DIoU}}^{(\mathcal{I}_{\hat{y}_c},\mathcal{I}_{y_c})}(n,m)\right) \quad (3)$$

$$\mathcal{L}_{\text{DIoU}}^{(\text{GT})} = \sum_{n}^{N} \mathcal{L}_{GDF}^{(\text{GT})}(n) \cdot \min_{m \in M}\left(\mathcal{L}_{\text{DIoU}}^{(\mathcal{I}_{\hat{y}_c},\mathcal{I}_{y_c})}(n,m)\right) \quad (4)$$

In our experiments, we compounded the Proximity loss with the GDF loss as suggested by the original paper [26], which is coded as GDF+Proximity in this paper and formalized as

$$\mathcal{L}_{GDF+\text{Proximity}} := \mathcal{L}_{GDF} + 0.1 \cdot \mathcal{L}_{\text{Proximity}} \quad (5)$$

where $\mathcal{L}_{GDF}$ represents the GDF loss and $\mathcal{L}_{\text{proximity}}$ represents the Proximity loss with a weight of 0.1.

E. Performance Metrics

We first measured the quality of the CMF fracture segmentation by using the Dice similarity coefficient (DSC) at the voxel level. To avoid a bias towards background voxels, we also measured the quality of CMF fracture segmentation and detection at the fracture/instance level by using DSC (DSC_{ins}), positive predictive value (PPV), sensitivity (SEN), false detection rate (FDR), false negative rate (FNR), and F1-score (F1S). For the fracture/instance level metrics, an instance of ground truth CMF fracture is a correct detection (i.e., true positive or TP) if any of its voxels are segmented and is a missed detection (i.e., false negative or FN) if none of its

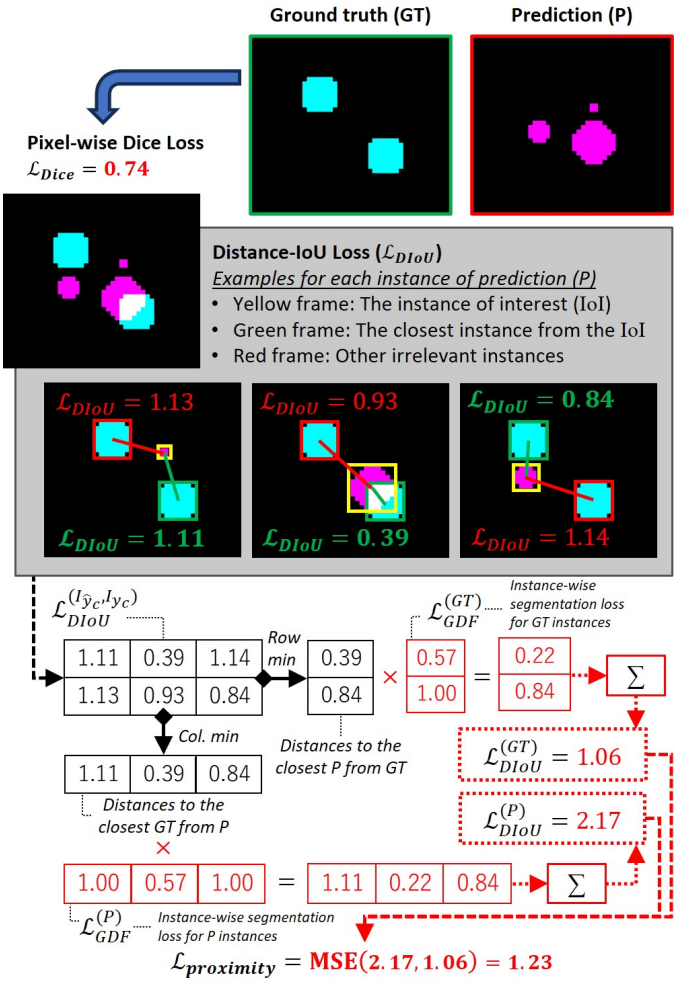

Fig. 3: Illustration of the use of Distance-IoU loss ($\mathcal{L}_{\mathrm{DIoU}}$) [30] in the Instance Proximity loss [26]. $\mathcal{L}_{\mathrm{DIoU}}$ is computed by using (1). Cyan indicates the ground truth instances, magenta indicates the predicted segmentation instances, the yellow frame indicates the instance of interest, the green frame indicates the closest instance from the instance of interest, and the red frame indicates other irrelevant instances. Everything illustrated in red color located outside the DIoU box means that the gradients are available for computation (see [26]).

voxels are correctly segmented. On the other hand, an instance of predicted CMF fracture is a false positive (FP) if it does not intersect with any ground truth CMF fracture instances.

Lastly, to determine which models produced the best overall results, a numeric rank (r) was assigned to each performance measurement, such that mean ranks could be calculated (i.e., Rank All or RA).

V. Results

A. Quantitative Results: Patch-based vs. Full-image Training/Inference Approaches

Tables II and III show all quantitative performance metric values produced by different settings of training using the Swin UNETR model for both 3D patch-based and 3D full-image training/inference, respectively. Both tables show that incorporating CT and labels of skull regions (CT+SR) into the Swin UNETR model as input produced better detection and segmentation results compared to using only CT or CT combined with the binary label of the skull (CT+S), where the RA values for CT+SR are always lower and better than the RA values for CT and CT+S. In addition, the experiments show that the models trained with 3D full images (results in Table III) produced better detection results than the models trained with 3D patch-based images (results in Table II). The best and second best performing models trained with 3D patch-based images (Table II) are the P3 and P6 with RA values of 2.3 and 2.6, respectively, and F1S values of 0.5121 and 0.4448, respectively. Similarly, the best and second best performing models trained with 3D full images (Table III) are the F3 and F6 with RA values of 1.3 and 2.3, respectively, and F1S values of 0.5902 and 0.5534, respectively. Note that P3 and F3 were both trained by using the GDF loss, while the P6 and F6 were both trained by using a compound loss of GDF loss and Proximity loss (GDF+Proximity).

B. Ensemble Inference Results

Although the best-performing models in Tables II and III produced relatively high F1S values, both P3 and F3 still missed many CMF fractures as indicated by relatively high FNR values (i.e., 0.3787 and 0.4031 for P3 and F3, respectively) and low PPV values (i.e., 0.4356 and 0.5837 for P3 and F3, respectively). The pixel-level and instance-level segmentation of CMF fractures (i.e., DSC and DSC_{ins}) were also noticeably low where the best values were produced by the P6 in Table II with DSC of 0.2995 and DSC_{ins} of 0.1561.

To further improve the performance, we tested an ensemble inference approach by combining two segmentation results from two different models, one from patch-based training/inference and one from full-image training/inference. Soft voting ensemble inference was performed where two probability values from two different segmentation models are averaged and then considered as part of the CMF fracture if the average value is ≥ 0.5. The models chosen for the ensemble inference are the best and second best performing models from both patch-based and full-image inference approaches, which are the P3, P6, F3, and F6 models.

Table IV shows that the ensemble inference approach improved the detection results of all tested ensemble models, where the P6+F3 model emerged as the best-performing model and produced the best SEN, FNR, and F1S values of 0.7708, 0.2292, and 0.6339, respectively. However, it is worth mentioning that the ensemble inference approach produced only slight or no improvement for DSC, DSC_{ins}, PPV, and FDR measurement metrics. This finding suggests that segmenting CMF fracture lines is a difficult task, as indicated by low DSC and DSC_{ins} values because they do not have clear borders that can be easily delineated even by an expert. In addition, the low PPV values and high FDR values indicate that CMF fractures are ambiguous due to the complex anatomical structure of the skull in which some normal structures are segmented and detected as CMF fractures (i.e., false positive detection).

TABLE II: Results for segmentation and detection of CMF fractures by performing 3D patch-based training and inference. Alphanumeric characters written in bold indicate the best values/rankings in each metric while the underlined ones indicate the second best values/rankings. Arrow symbols of ↑ and ↓ indicate that higher and lower values are better, respectively.

Code	Input	Loss	RA ↓	DSC ↑	r ↓	DSC_{ins} ↑	r ↓	PPV ↑	r ↓	SEN ↑	r ↓	FNR ↓	r ↓	FDR ↓	r ↓	F1S ↑	r ↓
P1	CT	GDF	6.0	0.0393	6	0.0139	6	0.2173	6	0.1866	6	0.8134	6	0.7827	6	0.2008	6
P2	CT+S	GDF	3.1	0.2813	3	0.1349	3	**0.6031**	**1**	0.3969	5	0.6210	5	0.6192	3	0.4788	2
P3	**CT+SR**	**GDF**	**2.3**	0.2773	4	0.1322	4	0.4356	2	0.6213	2	0.3787	2	**0.5644**	**1**	**0.5121**	**1**
P4	CT	GDF+Proximity	2.9	0.2946	2	0.1430	2	0.3842	3	0.5528	4	0.4472	4	0.6158	2	0.4533	3
P5	CT+S	GDF+Proximity	4.1	0.2462	5	0.1155	5	0.3415	4	0.5795	3	0.4205	3	0.6585	4	0.4298	5
P6	CT+SR	GDF+Proximity	2.6	**0.2995**	**1**	**0.1561**	**1**	0.3392	5	**0.6456**	**1**	**0.3544**	**1**	0.6608	5	0.4448	4

TABLE III: Results for segmentation and detection of CMF fractures by performing 3D full-image training and inference. Alphanumeric characters written in bold indicate the best values/rankings in each metric while the underlined ones indicate the second best values/rankings. Arrow symbols of ↑ and ↓ indicate that higher and lower values are better, respectively.

Code	Input	Loss	RA ↓	DSC ↑	r ↓	DSC_{ins} ↑	r ↓	PPV ↑	r ↓	SEN ↑	r ↓	FNR ↓	r ↓	FDR ↓	r ↓	F1S ↑	r ↓
F1	CT	GDF	5.7	0.1810	6	0.0592	6	0.5119	6	0.4416	5	0.5584	5	0.4881	6	0.4742	6
F2	CT+S	GDF	3.6	0.2104	3	0.0862	3	0.5579	4	0.4768	4	0.5232	4	0.4421	4	0.5142	3
F3	**CT+SR**	**GDF**	**1.3**	**0.2660**	**1**	**0.1067**	**1**	0.5837	2	**0.5969**	**1**	**0.4031**	**1**	0.4163	2	**0.5902**	**1**
F4	CT	GDF+Proximity	4.0	0.2040	4	0.0856	4	0.5223	5	0.4866	3	0.5134	3	0.4777	5	0.5038	4
F5	CT+S	GDF+Proximity	4.1	0.1970	5	0.0688	5	**0.6075**	**1**	0.4298	6	0.5702	6	**0.3925**	**1**	0.5034	5
F6	CT+SR	GDF+Proximity	2.3	0.2384	2	0.0926	2	0.5596	3	0.5473	2	0.4527	2	0.4404	3	0.5534	2

TABLE IV: Results for segmentation and detection of CMF fractures by performing ensemble inference using the best and second best models from patch-based and full-image training (i.e., the P3 and P6 from Table II and the F3 and F6 from Table III). Alphanumeric characters written in bold indicate the best values/rankings in each metric while the underlined ones indicate the second best values/rankings. Arrow symbols of ↑ and ↓ indicate that higher and lower values are better, respectively.

Code	Input	Loss	RA ↓	DSC ↑	r ↓	DSC_{ins} ↑	r ↓	PPV ↑	r ↓	SEN ↑	r ↓	FNR ↓	r ↓	FDR ↓	r ↓	F1S ↑	r ↓
P3+F3	CT+SR	GDF	3.7	0.3026	4	0.1234	4	0.5074	4	0.6948	3	0.3052	3	0.4926	4	0.5865	4
P6+F6	CT+SR	GDF+Proximity	2.6	0.3065	3	0.1273	3	0.5182	3	0.7669	2	0.2331	2	0.4818	3	0.6185	2
P3+F6	CT+SR	Mixed	5.4	0.2939	5	0.1192	5	0.4833	6	0.6820	5	0.3180	5	0.5167	6	0.5657	6
P6+F3	**CT+SR**	**Mixed**	**1.6**	0.3134	2	0.1285	2	0.5384	2	**0.7708**	**1**	**0.2292**	**1**	0.4616	2	**0.6339**	**1**
P3+P6	CT+SR	Mixed	3.6	**0.3206**	**1**	**0.1388**	**1**	0.5046	5	0.6938	4	0.3062	4	0.4954	5	0.5842	5
F3+F6	CT+SR	Mixed	4.1	0.2787	6	0.1137	6	**0.5516**	**1**	0.6601	6	0.3399	6	**0.4484**	**1**	0.6010	3

TABLE V: Clinical assessments of CMF fracture detection produced by the P1 model (i.e., the baseline model) and the P6+F3 model (i.e., the best performing model) on different regions of the skull. Alphanumeric characters written in bold indicate the best values in each metric. Arrow symbols of ↑ and ↓ indicate that higher and lower values are better, respectively. Dash symbol "-" indicates that the value cannot be calculated due to division by zero because there are no CMF fractures in the manual labels.

Region	P1 (baseline)					P6+F3 (best)				
	PPV ↑	TPR ↑	FDR ↓	FNR ↓	F1S ↑	PPV ↑	TPR ↑	FDR ↓	FNR ↓	F1S ↑
R1 - Mandible	0.5000	0.0909	0.5000	0.9091	0.1538	**0.6842**	**0.7222**	**0.3158**	**0.2778**	**0.7027**
R2 - Maxilla	0.6818	0.5172	0.3182	0.4828	0.5882	**0.7400**	**1.0000**	**0.2600**	**0.0000**	**0.8506**
R3 - Zygoma	0.6000	0.2143	0.4000	0.7857	0.3158	**0.6667**	**1.0000**	**0.3333**	**0.0000**	**0.8000**
R4 - Frontal Cranial Vault	0.3333	0.5000	0.6667	0.5000	0.4000	**0.7500**	**1.0000**	**0.2500**	**0.0000**	**0.8571**
R5 - Spheno & Temporo Lateral	0.0000	0.0000	1.0000	1.0000	0.0000	**0.9333**	**1.0000**	**0.0667**	**0.0000**	**0.9655**
R6 - Occipital Cranial Vault	0.0000	-	1.0000	-	-	**-**	**-**	**-**	**-**	**-**
R7 - Parietal Cranial Vault	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**
All Regions	0.5556	0.2985	0.4444	0.7015	0.3883	**0.7431**	**0.9419**	**0.2569**	**0.0581**	**0.8308**

C. Visual and Clinical Assessment Results

To validate the feasibility of using the proposed automated CMF fracture detection in the real world, we performed a clinical assessment in which experts visually assessed the results produced by the P1 (baseline) and P6+F3 (best) models. In contrast to the quantitative performance analysis, we assessed whether the predicted CMF fractures were located in the same skull region as the actual CMF fracture and ignored small spatial translations such as under-/over-segmentation (i.e., the predicted CMF fracture does not have to be in the exact location of the actual CMF fracture). This type of clinical assessment was primarily performed to explore whether the proposed approach could be used to pinpoint potential locations of CMF fractures in clinical conditions (e.g., screening). Examples of CMF fracture visualization used for clinical assessment are shown in Fig. 5. We tabulated the clinical assessment as PPV, TPR, FDR, FNR, and F1S values shown in Table V and visualized them in Fig. 4.

Table V and Fig. 4 show that the P1 (baseline) model missed many CMF fractures with high FNR values whereas the P6+F3 (best) model missed some CMF fractures only in the mandible region (R1). The P6+F3 model produced several false positive

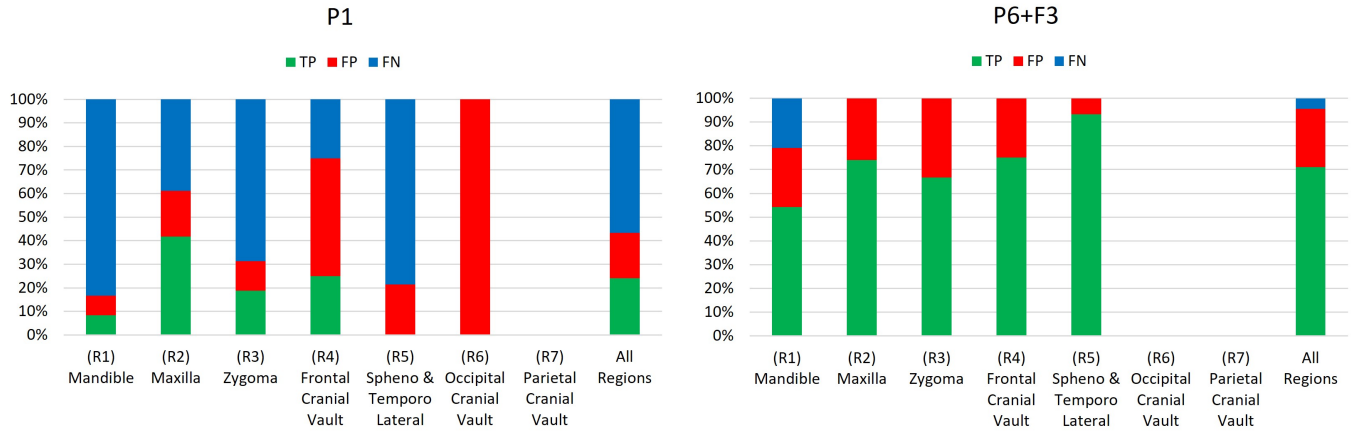

Fig. 4: Visualization of the tabulated clinical assessment results performed by experts in the forms of TP, FN, and FP values for each skull's regions produced by the P1 (baseline) and P6+F3 (best) models. This visualization is based on Table V.

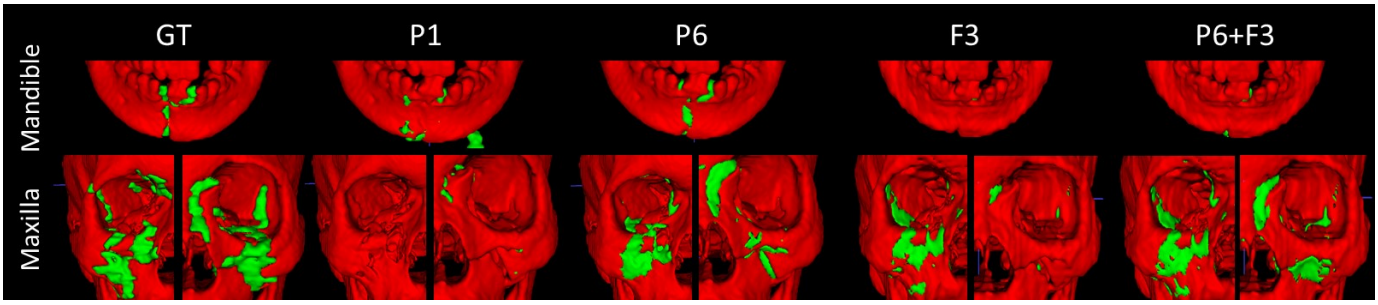

Fig. 5: Visualization of CMF fractures segmentation from three different subjects in the mandible and maxilla regions of the skull based on manual labeling or ground truth (GT) and different training settings of 3D Swin UNETR, which are the P1 (baseline), P6, F3, and P6+F3 (best) models. Red represents the skull area while green represents the CMF fractures. For each subject, only part of the skull is shown to maintain anonymity.

detections, but they are easily recognized as such by experts and are beneficial for pinpointing potential CMF fractures in the screening (i.e., faster screening process by experts). Thus, it is clear that the P6+F3 model performed much better in the clinical assessment than the P1 model with better PPV, TPR, FDR, FNR, and F1S values (i.e., 0.7431, 0.9419, 0.2569, 0.0581, and 0.8308, respectively) in all regions of the skull.

The experts also gave several important feedback about the proposed automated detection and segmentation of CMF fractures for further development, which are:

(1) the tested segmentation models often under-/over-segment the CMF fractures, which led to low DSC scores in the quantitative results,
(2) CMF fractures were often segmented as multiple fractures that did not link to each other and some of them did not intersect with the manual labels, which led to low PPV and high FDR values in the quantitative results,
(3) CMF fractures were often difficult to detect and segment due to various types of fractures (e.g., bone loss, displacement, fragmentation, etc.), which led to under-segmentation or no detections,
(4) several gaps in the skull were not labeled as CMF fractures by the experts due to various reasons (e.g., low quality of CT scan images that led to unintended omissions, joints or gaps between different bones instead of CMF fractures, and gaps were not classified as CMF fractures because experts considered other parameters such as the presence of hematoma), and
(5) CMF fractures present in subjects with many CMF fractures were more easily detected and segmented by the model compared to those in subjects with fewer and smaller CMF fractures.

VI. Conclusion

In this study, we developed an automated detection system for craniomaxillofacial (CMF) fractures from head CT images using ensembles of deep segmentation neural networks using the 3D Swin UNETR model. The combination of a state-of-the-art 3D Swin UNETR model, additional input in the form of labels of skull regions, a compound loss of Generalized Dice, Focal, and Proximity losses, and an ensemble inference approach successfully achieved good detection results in both quantitative and clinical assessments. Clinical evaluations were manually performed by multiple experts where the best-performing model achieved the PPV of 82.49%, TPR of 96.03%, FDR of 17.51%, FNR of 3.97%, and F1-score of 88.23%. However, the best-performing model often under-/over-segmented the CMF fractures because CMF fractures

have unclear boundaries that are even difficult for experts to delineate, resulting in low DSC scores. Furthermore, although the best-performing model still produced some false positive detections, they can easily be recognized as such by experts and the results are still beneficial for pinpointing potential CMF fractures and speeding up the diagnostic process in screening. These findings confirm the feasibility of an automated CMF fracture detection via segmentation but also highlight inherent limitations. Further assessment using prospective data is needed to demonstrate the real potential use in clinical settings. Furthermore, studies on the development of automated skull region segmentation are needed for a fully automated system.

REFERENCES

[1] N. A. Anggayanti, E. Sjamsudin, T. Maulina, and A. Iskandarsyah, "The quality of life in the treatment of maxillofacial fractures using open reduction: A prospective study," *Age*, vol. 2, pp. 13–33, 2020.

[2] C. A. Taylor, J. M. Bell, M. J. Breiding, and L. Xu, "Traumatic brain injury–related emergency department visits, hospitalizations, and deaths—united states, 2007 and 2013," *MMWR Surveillance Summaries*, vol. 66, no. 9, p. 1, 2017.

[3] S. Shah, S. K. Uppal, R. K. Mittal, R. Garg, and K. Saggar, "Diagnostic tools in maxillofacial fractures: Is there really a need of three-dimensional computed tomography?," *Indian Journal of Plastic Surgery*, vol. 49, no. 02, pp. 225–233, 2016.

[4] C. H. Buitrago-Téllez, C.-P. Cornelius, J. Prein, C. Kunz, A. Di Ieva, and L. Audigé, "The comprehensive aocmf classification system: radiological issues and systematic approach," *Craniomaxillofacial Trauma & Reconstruction*, vol. 7, no. 1_suppl, pp. 123–130, 2014.

[5] J. Mayer, D. Wainwright, J. Yeakley, K. Lee, J. Harris Jr, and M. Kulkarni, "The role of three-dimensional computed tomography in the management of maxillofacial trauma," *Journal of Trauma and Acute Care Surgery*, vol. 28, no. 7, pp. 1043–1053, 1988.

[6] K. Warin, W. Limprasert, S. Suebnukarn, T. Paipongna, P. Jantana, and S. Vicharueang, "Maxillofacial fracture detection and classification in computed tomography images using convolutional neural network-based models," *Scientific Reports*, vol. 13, no. 1, p. 3434, 2023.

[7] Y. Tong, B. Jie, X. Wang, Z. Xu, P. Ding, and Y. He, "Is convolutional neural network accurate for automatic detection of zygomatic fractures on computed tomography?," *Journal of Oral and Maxillofacial Surgery*, 2023.

[8] A. P. Brady, "Error and discrepancy in radiology: inevitable or avoidable?," *Insights into imaging*, vol. 8, pp. 171–182, 2017.

[9] G. Yuan, G. Liu, X. Wu, and R. Jiang, "An improved yolov5 for skull fracture detection," in *International Symposium on Intelligence Computation and Applications*, pp. 175–188, Springer, 2021.

[10] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov, "Applications of deep learning in biomedicine," *Molecular pharmaceutics*, vol. 13, no. 5, pp. 1445–1454, 2016.

[11] W. M. D. Wan Zaki, M. F. Ahmad Fauzi, and R. Besar, "A new approach of skull fracture detection in ct brain images," in *International Visual Informatics Conference*, pp. 156–167, Springer, 2009.

[12] A. Yamada, A. Teramoto, T. Otsuka, K. Kudo, H. Anno, and H. Fujita, "Preliminary study on the automated skull fracture detection in ct images using black-hat transform," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6437–6440, IEEE, 2016.

[13] C.-Y. Yang, C.-H. Lo, H.-C. Wang, J.-H. Chou, and Y.-C. F. Wang, "Weakly-supervised learning for attention-guided skull fracture classification in computed tomography imaging," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1337–1341, IEEE, 2019.

[14] X. Wang, Z. Xu, Y. Tong, L. Xia, B. Jie, P. Ding, H. Bai, Y. Zhang, and Y. He, "Detection and classification of mandibular fracture on ct scan using deep convolutional neural network," *Clinical Oral Investigations*, pp. 1–9, 2022.

[15] Y. J. Seol, Y. J. Kim, Y. S. Kim, Y. W. Cheon, and K. G. Kim, "A study on 3d deep learning-based automatic diagnosis of nasal fractures," *Sensors*, vol. 22, no. 2, p. 506, 2022.

[16] M. Amodeo, V. Abbate, P. Arpaia, R. Cuocolo, G. Dell'Aversana Orabona, M. Murero, M. Parvis, R. Prevete, and L. Ugga, "Transfer learning for an automated detection system of fractures in patients with maxillofacial trauma," *Applied Sciences*, vol. 11, no. 14, p. 6293, 2021.

[17] G. Liu, Q. Wu, G. Yuan, and X. Wu, "Skull fracture detection method based on improved feature pyramid network," in *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)*, pp. 756–762, IEEE, 2021.

[18] Z. Kuang, X. Deng, L. Yu, H. Zhang, X. Lin, and H. Ma, "Skull r-cnn: A cnn-based network for the skull fracture detection," in *Medical Imaging with Deep Learning*, pp. 382–392, PMLR, 2020.

[19] X. Lin, Z. Yan, Z. Kuang, H. Zhang, X. Deng, and L. Yu, "Fracture r-cnn: An anchor-efficient anti-interference framework for skull fracture detection in ct images," *Medical Physics*, 2022.

[20] W. Shan, J. Guo, X. Mao, Y. Zhang, Y. Huang, S. Wang, Z. Li, X. Meng, P. Zhang, Z. Wu, *et al.*, "Automated identification of skull fractures with deep learning: a comparison between object detection and segmentation approach," *Frontiers in Neurology*, vol. 12, 2021.

[21] T. S. Jeong, G. T. Yee, K. G. Kim, Y. J. Kim, S. G. Lee, and W. K. Kim, "Automatically diagnosing skull fractures using an object detection method and deep learning algorithm in plain radiography images," *Journal of Korean Neurosurgical Society*, 2022.

[22] M. F. Rachmadi, M. d. C. Valdes-Hernandez, M. L. F. Agan, C. Di Perri, T. Komura, A. D. N. Initiative, *et al.*, "Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain mri with none or mild vascular pathology," *Computerized Medical Imaging and Graphics*, vol. 66, pp. 28–43, 2018.

[23] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pp. 272–284, Springer, 2022.

[24] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 240–248, Springer, 2017.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[26] M. F. Rachmadi, M. Byra, and H. Skibbe, "A new family of instance-level loss functions for improving instance-level segmentation and detection of white matter hyperintensities in routine clinical brain mri," *Computers in Biology and Medicine*, p. 108414, 2024.

[27] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murray, A. Myronenko, C. Zhao, D. Yang, V. Nath, Y. He, Z. Xu, A. Hatamizadeh, W. Zhu, Y. Liu, M. Zheng, Y. Tang, I. Yang, M. Zephyr, B. Hashemian, S. Alle, M. Zalbagi Darestani, C. Budd, M. Modat, T. Vercauteren, G. Wang, Y. Li, Y. Hu, Y. Fu, B. Gorman, H. Johnson, B. Genereaux, B. S. Erdal, V. Gupta, A. Diaz-Pinto, A. Dourson, L. Maier-Hein, P. F. Jaeger, M. Baumgartner, J. Kalpathy-Cramer, M. Flores, J. Kirby, L. A. Cooper, H. R. Roth, D. Xu, D. Bericat, R. Floca, S. K. Zhou, H. Shuaib, K. Farahani, K. H. Maier-Hein, S. Aylward, P. Dogra, S. Ourselin, and A. Feng, "MONAI: An open-source framework for deep learning in healthcare," Nov. 2022.

[28] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, *et al.*, "U-net: deep learning for cell counting, detection, and morphometry," *Nature methods*, vol. 16, no. 1, pp. 67–70, 2019.

[29] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, IEEE, 2016.

[30] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 12993–13000, 2020.