IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 March 2024, accepted 14 April 2024, date of publication 24 April 2024, date of current version 6 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3393231

RESEARCH ARTICLE

Two-Stage Approach to Intracranial Hemorrhage Segmentation From Head CT Images

JAGATH C. RAJAPAKSE[1], (Fellow, IEEE), CHUN HUNG HOW[1], YI HAO CHAN[1], LUKE CHIN PENG HAO[1], ABHINANDAN PADHI[1], VIVEK ADRAKATTI[1], IRAM RAIS ALAM KHAN[2], AND TCHOYOSON LIM[2,3]

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798
[2]National Neuroscience Institute, Singapore 308433
[3]Duke-NUS Medical School, Singapore 169857

Corresponding author: Jagath C. Rajapakse (asjagath@ntu.edu.sg)

ABSTRACT Intracranial hemorrhage (ICH) is an emergency and a potentially life-threatening condition. Automated segmentation of ICH from head CT images can provide clinicians with volumetric measures that can be used for diagnosis and decision support for treatment procedures. Existing solutions typically involve training deep learning models to perform segmentation directly on the whole CT image. However, datasets with segmentation masks are typically very small in comparison with datasets with bounding boxes. Thus, we propose a two-stage approach that utilizes both bounding boxes and segmentation masks to help improve segmentation performance. In the first stage, ICH regions are detected and localized with bounding boxes surrounding the lesion by using a supervised YOLOv5 object detector. In the second stage, the localized ICH foreground is automatically segmented using TransDeepLab, an attention-based transformer network. Although we utilize both ground-truth bounding boxes and segmentation masks, different datasets can be used to train each stage. There is no requirement for pairing up bounding boxes and segmentation masks to train the model. Since bounding box annotations are available in larger quantities than segmentation masks, our approach allows these large datasets of bounding boxes to be used to improve ICH segmentation performance. On our dataset of segmentation masks, we demonstrated that our proposed two-stage YOLOv5 + TransDeepLab model outperformed segmentation methods such as SegResNet by 8% in terms of Dice score. Given ground truth bounding boxes, a Dice score of 0.769 is achieved, outperforming state-of-the-art methods such as nnU-Net. In sum, our proposed two-stage approach produces more accurate binary segmentation of ICH for neuroradiologists and these improved measurements could potentially aid their clinical decision-making process.

INDEX TERMS Brain lesion segmentation, DeepLab, intracranial hemorrhage, object detection, YOLO.

I. INTRODUCTION

Intracranial Hemorrhage (ICH) refers to extravascular accumulation of blood within intracranial spaces. The causes of ICH are diverse, including head trauma, hypertensive hemorrhage, vascular malformations, tumors, cerebral venous thrombosis, and cerebral amyloid angiopathy, among other causes [1]. In serious cases, it can lead to permanent neurological damage or even death. ICH is a life-threatening condition as the 30-day mortality rate for ICH ranges from 35% to 52%, where only 20% of survivors are expected to have full functional recovery within 6 months [4]. It is also quite a common condition. The frequency of acute ICH worldwide is 24.6 per 100,000 persons per year, with approximately 40,000 to 67,000 cases per year in the United States [4]. Thus, there is a need for solutions that can detect ICH to allow for timely intervention by clinicians.

Non-contrast computed tomography (CT) is the first line of imaging for neurological diseases. It is also the main modality used under emergency conditions (such as ICH) due to its accessibility and speed [2]. To accurately locate and identify the type of ICH, head CT images are examined

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

by neuroradiologists. This process can be challenging and time-consuming [3]. Acute blood appears hyperdense (white) on CT, posing little difficulty in diagnosis. However, ICH detection and interpretation may pose some challenges to junior radiologists, especially if the amount of bleed is small. The development of automated detection and segmentation techniques can aid radiologists in their work [4] and potentially reduce misinterpretation of some common ICH subtypes, especially by trainee radiologists [5]. Having a fast and comprehensive algorithm for ICH segmentation could reduce patient turn-around time and expedite the diagnosis of ICH.

Existing methods rely on traditional statistical modeling approaches or deep neural networks for semantic segmentation. Traditional statistical modeling approaches such as Expectation Maximization (EM) algorithm are unsupervised and do not require training data. However, they do not usually achieve high segmentation performance because they may erroneously mark non-ICH regions (such as the skull, calcifications, noise, artifacts, etc.) as ICH. Karkkainen et al. [6] proposed to run EM algorithm on each CT image voxel to detect voxels that contain ICH, but doing so requires the optimization algorithm to be applied iteratively on every voxel, which is computationally expensive.

Supervised deep neural networks for segmentation like Fully Convolutional Networks (FCN) [7] are generally more accurate than unsupervised approaches. For medical image segmentation tasks, the U-Net architecture [8] has been very popular due to its consistent and outstanding performance. Lately, inspired by the original transformer architecture used in natural language processing [9], Vision Transformer (ViT) [10] has been introduced for medical image analysis to capture long-range dependencies that are missed out by U-Net [11]. Hybrid architectures that combine U-Net and ViT have also been proposed. For example, Swin-Unet [12] has a similar structure as the classic U-Net (down-sampling encoder, up-sampling decoder, skip-connections, and bottleneck layer), but the convolutional blocks are replaced by Swin Transformer blocks [13] and the patch merging module performs down-sampling. TransDeepLab [14], an extension of DeepLab [15], uses Swin Transformer blocks to encode the image and the bottleneck performs Atrous Spatial Pyramid Pooling to exploit multi-scale features from the hierarchical encoder.

Variations of the above-mentioned deep learning models have been adapted to solve domain-specific challenges faced when segmenting ICH lesions. For example, a CT scan comprises a sequence of slices. Introducing RNN and LSTM can help to capture dependencies across image slices. Redman et al. proposed a DenseNet + Long-Short Term Memory (LSTM) approach to perform classification and segmentation using sequential 2D slices [5]. They focused only on binary classification (presence vs absence of ICH) as the main task and segmented hemorrhagic regions as auxiliary tasks. Similarly, Ye et al. [16] proposed a combined model of convolutional neural networks and recurrent neural networks (CNN + RNN) to predict ICH subtypes. As it is also helpful for clinicians to know the volume of the bleeds for disease prognosis, Kuo et al. proposed PatchFCN as an end-to-end network to perform segmentation for each image patch, resembling how radiologists would carefully study a specific region in a slice [17]. Another domain-specific challenge is the issue of data scarcity, which is tackled by AMD-DAS [18] by using data from other similar domains such as magnetic resonance imaging (MRI) images of brain tumors. They trained a generative model to generate pseudo-CT images from MRI images that contain segmented glioma. These generated images help to augment the original dataset. While the target CT image has no ground truth mask, the model is trained to segment in pseudo-CT space for the task of Intraparenchymal hemorrhage (IPH) segmentation. Finally, Zhao et al. [19] proposed to use no-new-Net (nnU-Net) to quantify the volume of ICH, Intraventricular hemorrhage (IVH) and peripheral edema in 3D. nnU-Net [20] is based on the U-Net architecture but adds on automatic configuration of key design choices of a typical image segmentation pipeline, such as preprocessing, augmentation and network architecture.

Despite these advancements, several challenges remain in ICH segmentation. Most existing architectures still require sizable training data that is infeasible to collect and label. This is especially the case for segmentation masks of ICH lesions, which requires expert knowledge from neuroradiologists. In the case of AMD-DAS [18], segmentation masks are still required even in their weak supervision approach. Another important consideration for ICH segmentation is the runtime of the algorithm, which cannot be too long since ICH is an emergency. For instance, EM takes 10 times longer than most deep learning approaches (based on our experiments shown below in Table 3), making it less suited for clinical deployment despite its strength of not requiring labelled data.

In this study, we propose a novel technique for ICH segmentation on head CT slices which involves two stages. The first stage involves training a YOLOv5 model to detect ICH lesions and draw bounding boxes around them. The second stage involves training a TransDeepLab model to segment regions with ICH within each bounding box produced from the first stage. Such a two-stage pipeline has two main benefits: (1) it leverages the ease of creating ground truth labels for object detection relative to semantic segmentation, and (2) the ICH localization prior to segmentation helps to reduce background noise. Manual annotation of bounding boxes only involves identifying their coordinates. This is much easier and quicker than preparing segmentation masks. Thus, datasets with bounding boxes (e.g. the ''Brain Hemorrhage Extended'' dataset has annotations for almost 40,000 slices) are available in larger quantities than datasets with segmentation masks (if available, usually at most in the hundreds and much fewer for complex and irregular lesions). Our approach leverages these large bounding box datasets to maximize segmentation performance. Although a two-stage approach

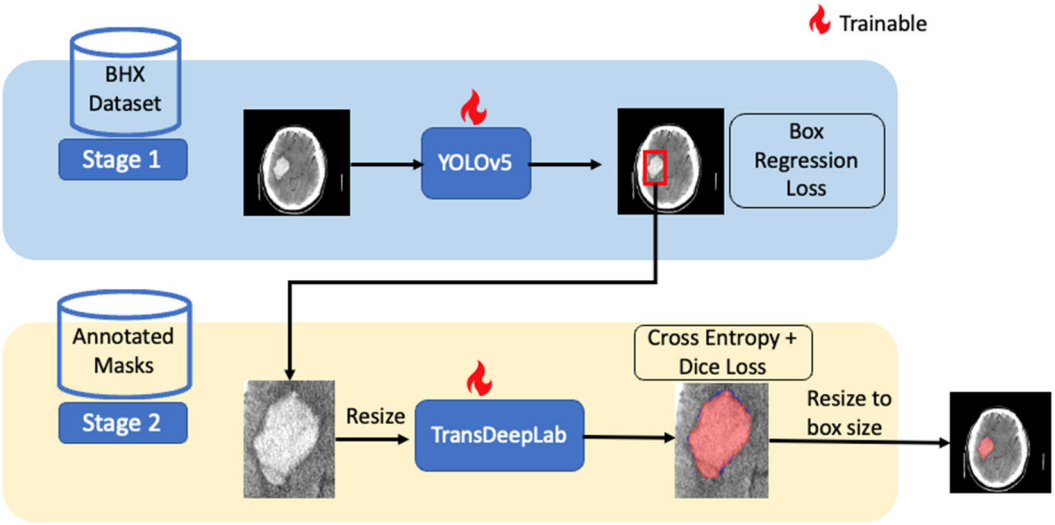


FIGURE 1. Illustration of our two-stage approach.

seems more complex, the runtime of our approach is similar to single-stage approaches, in part due to the highly optimized YOLO architecture.

Two-stage approaches have been successfully used in other domains. Yi et al. [21] proposed a framework with two branches: detection and segmentation. Object detection helps the segmentation branch to focus within the region of interest by cropping region of interest patches. They introduced skip connections between the detection and segmentation branch and performed instance normalization with learned statistics within the bounding box. This removes the statistics of neighboring objects and recovers morphological details of the main object to be segmented. In contrast, we trained a segmentation model to accurately segment the lesion with complex shapes within the detected bounding box. Other closely related works include architectures that are capable of simultaneous detection and segmentation. Bhattacharya implemented Mask RCNN on ICH segmentation, which typically requires both ground-truth bounding boxes as well as segmentation masks to be provided simultaneously during training [22]. On the other hand, our approach allows the detection and segmentation stages to be trained independently with different datasets. In addition, our two-stage approach is more flexible as it allows the use of any combination of detection model and segmentation model, while Mask RCNN presents as a single architecture used for both tasks and changing parts of the architecture could present significant technical challenges in terms of code implementation.

Additionally, while YOLO was mostly used for object localization, Jiang et al. [23] proposed to use YOLO to localize the 6 key points for each tooth in the panoramic film to determine the stage of periodontal bone loss. To reduce the interference of various structures around the tooth, U-Net was used to obtain the contour of each tooth before the localization step. Our method shares the idea of zooming into the region of interest, but we used YOLO to localize the lesion before segmentation. Bai et al. [24] proposed YUSEG which consists of an ensemble of YOLOv5 models as the first stage to obtain the bounding boxes, and an Eff-UNet as the model used in the second stage to perform cell instance segmentation for each detected box. This is similar to our proposed approach, but do we not use an ensemble of detection models. Furthermore, in this paper, we conducted a more thorough study of various combinations of detection and segmentation models to find the optimal combination.

Overall, the key contributions of our work are: (i) a two-stage approach of ICH segmentation that leverages the abundance of bounding box annotations to improve downstream segmentation performance, (ii) our experiment results demonstrated the superiority of two-stage approaches over single-stage approaches while having similar runtime, (iii) the finding that out of the various detection and segmentation models tested, YOLOv5+TransDeepLab was the best combination for ICH lesion segmentation.

II. METHOD

Figure 1 provides an overview of how training is done in our proposed two-stage approach. Given a CT image, lesions are detected with bounding boxes in the first stage via YOLOv5 (which is finetuned on BHX, a dataset with bounding boxes of ICH lesions). During the second stage, the weights in YOLOv5 are frozen. Lesions are segmented by processing only the parts of the images within bounding boxes and the TransDeepLab model is trained using a separate dataset containing segmentation masks only. Note that all the segmentation models in the experiments including our second stage model only perform binary segmentation.

A. DATA PREPROCESSING

The proposed pipeline accepts head CT slices in the Digital Imaging and Communications in Medicine (DICOM) format, which is the convention for medical data. These DICOM slices are converted to 512 × 512 pixel arrays. Pixel intensities in these arrays are represented in Hounsfield Units (HU). HU is a dimensionless unit used in CT images and its scale ranges from approximately −1024 to +3071 HU where −1000 HU represents the density of air and +2000 represents dense bone tissue. However, the default HU range usually provides poor contrast between normal brain tissue and ICH tissue. Hence, the contrast of each CT slice is adjusted through a process known as Windowing (Contrast Stretching) [25]. In this process, the window width and window level of each CT slice are adjusted. For the ICH detector, window width and level values were obtained from the DICOM metadata. After windowing, the pixel intensities are scaled to the range of 0 to 255.

In our implementation of the Expectation Maximization (EM) method, the CT slices underwent skull and calcification removal. Skull tissue and calcification in the brain appear as bright white regions on CT images due to their high density, even though both are normal and benign tissues respectively. Acute ICH regions also appear white on a CT image, albeit being slightly dimmer. Consequently, if an ICH lesion is located adjacent to the skull and/or benign calcification is present in a CT slice, the detection and segmentation algorithm may output false positives for the skull / calcification pixels. To address this problem, we removed pixels with intensity greater than 250 for each CT slice, which removes most skull / calcification pixels. At this point, the CT pixel arrays are saved as PNG files to speed up data loading for ICH detector at runtime. However, such preprocessing is not necessary for other supervised segmentation models.

B. DATA AUGMENTATION

We incorporated various types of data augmentation to train our models robustly and increase the size of the training dataset. When training the detection model, the following data augmentations were applied: augmentation of (hue, saturation, value) properties, translation, scaling, horizontal flips and creation of image mosaics. During the training of the segmentation model, random flipping and random 90° rotation were applied.

C. LESION DETECTION

In the first stage of the pipeline, lesion detection is performed by marking them out with bounding boxes. You Only Look Once (YOLO) is a family of state-of-the-art network architectures for the task of object detection [26]. There have been multiple iterations of the YOLO architecture, with YOLOv5 being the previous state-of-the-art version that is most widely used. YOLOv5 has 5 pre-trained networks to choose from [27] and [28] each with a different number of trainable parameters (e.g., YOLOv5n and YOLOv5x). The choice of pre-trained YOLOv5 network is made based on hardware constraints and the desired trade-off between detection speed and accuracy. Based on the hardware constraints and the high accuracy required for our use case, the YOLOv5l network (i.e., the second largest YOLOv5 pre-trained network) was chosen for our task. To choose the YOLOv5 hyperparameters that have the greatest influence on ICH detection accuracy, a Genetic Algorithm (GA) was used. We chose GA since other methods like grid search are infeasible due to the very large search spaces associated with YOLOv5's set of 29 hyperparameters. To perform lesion detection on head CT images, the 512 × 512 pre-processed CT images are provided as inputs to the YOLOv5 model. This produces bounding boxes that are subsequently used in the second stage.

D. LESION SEGMENTATION

In the second stage of the pipeline, we trained TransDeepLab [14], one of the state-of-the-art hybrid U-Net+Transformer models for segmentation [12]. Instance segmentation is performed within the bounding box, classifying the pixel as background or foreground, where the foreground represents the lesion. We fine-tuned the segmentation model on our dataset (dataset details presented in the next section) by adhering to the original TransDeepLab training configuration, such as resizing input image size to 224 × 224×3) via bicubic interpolation and duplicating the third channel 3 times. The output channel of the last convolutional layer of TransDeepLab is set to 2.

It is important to note that the source of bounding boxes for use in the second stage is different during training and inference. During model training of the second stage, since most datasets of annotation masks did not contain bounding box coordinates, instead we obtained the bounding boxes generated from the ground truth lesion mask using the mask_to_boxes algorithm from the PyTorch library. Then, we cropped the image region within the detected bounding boxes and subsequently resized the cropped region to 224 × 224 before segmentation was performed by TransDeepLab. The segmentation mask produced from the second stage is then resized to the size of the original bounding box. Similar steps were also performed during model inference, but the predicted bounding box from the first stage is used since no ground truth is accessible. In the scenario where multiple bounding boxes are predicted by the first stage, they are combined into one by creating the smallest possible bounding box that would encompass all predicted boxes. This new bounding box is then used to crop the image and follows the same resizing steps as described above.

Overall, our proposed two-stage approach involves using YOLOv5l in the first stage for lesion detection and TransDeepLab in the second stage for lesion segmentation. Each stage is trained independently. Performing lesion detection in the first stage helps the second stage to focus on the object of interest. To maximize the accuracy, we ensured via manual inspection that the ground truth bounding box fully covers

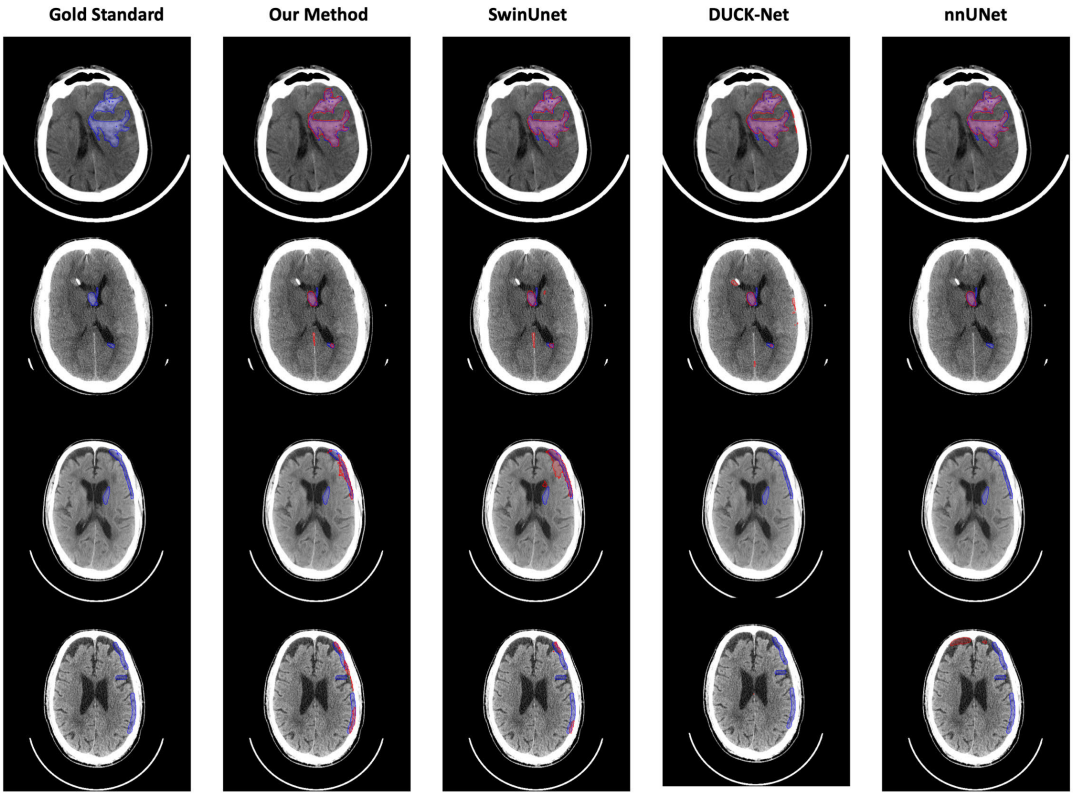


FIGURE 2. Examples of ground truth segmentation masks visualized along with the segmentation masks produced by our proposed two-stage ICH segmentation pipeline and other state-of-the-art segmentation models. Ground truth masks are shown in blue, and the predicted masks produced by various segmentation models are shown in red, overlaid on the ground truth segmentation masks.

the lesion segmented during training. Also, we note that in our implementation, the ground truth bounding boxes only contain a single subtype of lesion per box.

III. EXPERIMENTS AND RESULTS

In this section, we provide details about the datasets used, experiment setup, and results of our experiments. We then compare the performance of our pipeline to existing ICH segmentation techniques.

A. TRAINING DATASETS

Two data sources were used in this study: Qure.ai CQ500 dataset [29] and the Radiological Society of North America (RSNA) Intracranial Hemorrhage dataset [30]. The CQ500 dataset contains 491 head CT scans, with a total of 193,317 CT slices, while the RSNA dataset contains 874,035 head CT slices. Both datasets contain head CT slices of various subtypes of ICH. However, the CT images are only annotated with expert radiologists' reads regarding the presence of ICH in each CT slice and its subtype. They do not provide segmentation masks of the ICH lesions.

To train the detection model in the first stage, we used annotations from the Brain Hemorrhage Extended (BHX) dataset [31]. BHX is an extension of the CQ500 dataset that provides 39,668 ground-truth bounding box annotations for five types of acute ICH (subdural, epidural, subarachnoid, intraparenchymal, intraventricular) that are present in a total of 23,409 CQ500 CT slices. We used the BHX bounding boxes and their corresponding CT slices in CQ500 for ICH detection experiments using YOLOv5, with an 80:10:10 split for the training, validation, and test sets.

In addition, due to the paucity of openly accessible ICH segmentation masks, we manually curated a dataset with the supervision of a team of neurologists from our local hospital. A total of 347 randomly selected CT slices were taken from both datasets (approximately 70 slices for each of the five ICH types). Specifically, we annotated 100 CQ500 CT slices and 247 RSNA CT slices. These annotations include both ground-truth bounding boxes as well as segmentation masks for ICH tissues in the CT slices. The annotations were completed with the guidance of doctors and radiologists from the National Neuroscience Institute (NNI), Singapore [32]. In our experiments, we used 279 slices for training, 32 for validation and 36 for testing. Note that all slices contain lesions. Examples of these segmentation maps are shown in blue in Figure 2 (under the column 'Gold Standard').

B. EVALUATION METRICS

To evaluate the performance of the first stage in the pipeline (object detection tasks), mean Average Precision (mAP) was

computed at an Intersection over Union (IOU) threshold of 0.5, more commonly known as mAP@0.5. The performance of our segmentation stage was evaluated using two standard segmentation metrics, namely the IOU score and the Dice coefficient, also known as F1 score. Specifically, our method first performs segmentation on the image in the bounding box, which needs to be resized to the model input size. Then, the predicted mask was resized back to the size of the original box and overlaid on the original mask according to the box location. The Dice coefficient and IOU score at the second stage are computed using the final predicted mask and the original ground truth mask of each slice.

C. EXPERIMENT SETUP

Our two-stage pipeline involves a combined architecture of YOLOv5 with TransDeepLab. To justify this selection, we conducted experiments on alternative single-stage methods and experimented with various model combinations for two-stage methods. For single-stage methods, we compared against both unsupervised and supervised approaches. EM is an unsupervised algorithm, while DUCK-Net [33], SegResNet [34], SwinUnet [12], U-Net [8] and Feature Pyramid Network (FPN) [35] are deep, supervised segmentation network architectures. For two-stage methods, the use of a two-stage detection and segmentation approach allows for many combinations of architectures, and it is not clear which will be the best for ICH segmentation. Thus, we experimented with two widely used detection models (YOLOv5 and Faster RCNN) and three segmentation models (EM, SwinUnet and TransDeepLab). It is important to note again that for these two-stage models, inputs to the segmentation models are parts of the images that were within the bounding box, instead of using the whole image.

Our YOLOv5 object detection network was trained on bounding box coordinates and ICH subtype labels for 200 epochs with batch size 64, and with 29 hyperparameters as determined by a Genetic Algorithm (as explained in Section III). We obtained optimal results with YOLOv5 using an IOU threshold of 0.6 for non-max suppression. YOLOv5 obtained the following results on the test set: mAP@0.5 = 0.974, mAP@0.5:0.95 = 0.794. For the second stage, our TransDeepLab model was trained for 300 epochs with a learning rate of 0.01 and batch size of 8 using stochastic gradient descent (SGD) decaying optimizer. Our training method was stated in Section II-D.

For the implementation of the other single-stage models, we followed their original training configurations (such as optimizer and loss function), except that we reduced their number of parameters such that it became closer to our model in terms of size. This helps to ensure a fair comparison. For instance, we trained nnU-Net with its default data augmentation such as foreground oversampling but reduced the base number of features to 24. For other models such as DUCK-Net, SegResNet and SwinUNet, we only performed rotation and flipping as data augmentation. These models were trained with 300 epochs on the same training dataset.

D. COMPARING SINGLE-STAGE AND DOUBLE-STAGE METHODS

Table 1 shows the segmentation performance comparison between our method and alternative methods (single-stage and two-stage methods). From Table 1, it is evident that directly using an unsupervised approach such as EM produces very low Dice scores. However, when EM was used in a two-stage approach (e.g. Faster RCNN + EM, YOLOv5 + EM), the Dice score was much higher (0.557) and even outperformed most supervised single-stage models (e.g. U-Net [8] and more recent models like DUCK-Net [33]).

TABLE 1. Lesion segmentation performance for our method and state-of-the-art methods.

	Method	Dice score (F1 score)	Intersection over union (IOU)
Two-stage	**Ground truth boxes + TransDeepLab**	**0.769**	**0.657**
	YOLOv5 + TransDeepLab	**0.605**	**0.478**
	YOLOv5 + SwinUnet	0.597	0.473
	YOLOv5 + EM	0.557	0.436
	Faster RCNN + EM	0.480	0.356
Single-stage	nnU-Net [20]	0.665	0.566
	DUCK-Net [33]	0.523	0.416
	SegResNet [34]	0.522	0.411
	SwinUnet [12]	0.513	0.472
	U-Net [8]	0.458	0.351
	FPN [35]	0.390	0.332
	EM [6]	0.197	-

When the second stage is replaced by supervised models (i.e. combination of YOLOv5 and TransDeepLab), the Dice score improved by around 5% to 0.605. While well-optimized supervised approaches such as nnU-Net [20] obtained a Dice score of 0.665, we note that an optimal setup of our two-stage pipeline (i.e. predicted bounding boxes matches the ground truth) was able to achieve the best segmentation performance in terms of both evaluation metrics (Dice score of 0.769, IOU of 0.657). Further analysis of the confusion matrix revealed that this improvement over nnU-Net is driven primarily by higher true positives and lower false negatives. Additionally, further experiments done to compare these two models in a 5-fold cross-validation setting reveal that the improvement is statistically significant, with a p-value of 2.3×10^{-4} (two-sample t-test).

Overall, the results in Table 1 demonstrate the superiority of two-stage segmentation approaches over single-stage methods. These results provide empirical evidence demonstrating the value of using bounding boxes to narrow the field of view of the segmentation model, helping it to improve segmentation performance.

E. ABLATION STUDIES

Since YOLOv5 + TransDeepLab was the best-performing combination of models, an ablation study was performed to better understand the contributions of each stage of our proposed pipeline. We compared our proposed two-stage model to a direct approach of image segmentation using TransDeepLab. Note that in the two-stage model, while the model in the first stage performs object classification of lesion subtype, the second stage (segmentation model) was only trained on a binary segmentation task (presence / absence of ICH). For this ablation study, a separate TransDeepLab model was trained to segment ICH lesions regardless of the lesion subtype in the image and without any guidance from the bounding box. This segmentation model was trained for 600 epochs with SGD decaying optimizer. In Table 2, we observed that even after training the new segmentation model for twice as many epochs as the other segmentation models (which used 300 epochs), the segmentation performance of the single-stage TransDeepLab model was still lower than our method.

TABLE 2. Lesion segmentation performance for our 2-stage method and direct image segmentation.

	Method	Dice score (F1 score)	Intersection over union (IOU)
Two-stage	**Ground truth boxes + TransDeepLab**	**0.769**	**0.657**
	YOLOv5 + TransDeepLab	**0.605**	**0.478**
Single-stage	TransDeepLab [14]	0.566	0.446
	YOLOv5-Seg	0.480	0.380

We also trained a segmentation head using the body of the trained YOLOv5. During the conduct of our experiments, the authors of YOLOv5 had not implemented its segmentation head. Hence, we trained a segmentation head for YOLOv5, referred to as YOLOv5-seg in Table 2. We used a combination of 4 convolutional layers, 3 up-sampling layers and 2 Bottleneck Cross Stage Partial modules to up-sample the latent vector to the segmentation mask. As compared to the current official implementation of the segmentation head (Proto module) [27], we have more layers in our version of the model: we implemented two blocks, each containing a convolutional layer, up-sampling layer, and the YOLO C3 dense block. Between these two blocks, we also had two convolutional layers and an up-sampling layer. We trained the end-to-end YOLOv5-Seg model on the union of training and validation set until convergence is reached, with a batch size of 3 and stepped learning rate from 0.001. Our results in Table 2 demonstrated that our proposed two-stage method is more performant than the direct use of TransDeepLab as the segmentation model.

Overall, it is evident from the ablation study that TransDeepLab performs better than YOLOv5 with a segmentation head attached, possibly suggesting that the second stage has a larger contribution to the overall performance. However, our proposed approach of using the bounding boxes to narrow the field of view exposed to the segmentation model helps to further boost the performance. When ground truth bounding boxes are used, this improves segmentation performance by over 20% as compared to a single-stage TransDeepLab model, clearly showing the value of our two-stage pipeline.

F. COMPUTATIONAL COMPLEXITY

Since ICH is an emergency, it is important to ensure that the time taken to perform model inference is not too long. To record the time taken by each model to produce its predictions, we recorded the time before and after the model performs an inference of each image in the test set. This excludes the data preprocessing time of the model. The average inference time and standard deviation for each model are reported in Table 3. All experiments were performed on a server equipped with an Nvidia A100 GPU (PCIe 4.0, 80GB HBM2).

TABLE 3. Inference runtime of each model and their standard deviation.

	Method	Prediction time (seconds)
Two-stage	**YOLOv5 + TransDeepLab (Proposed method)**	0.52 ± 2.49
	YOLOv5 + SwinUnet	0.31 ± 1.98
	YOLOv5 + EM	3.75 ± 4.56
Single-stage	DUCK-Net [33]	0.48 ± 2.03
	SegResNet [34]	0.34 ± 1.89
	SwinUnet [12]	0.24 ± 1.98
	U-Net [8]	0.28 ± 1.31
	EM [6]	3.68 ± 4.56
	TransDeepLab [14]	0.45 ± 2.49

From Table 3, it is evident that most models can perform segmentation in around 0.3-0.5 seconds. This amounts to approximately 15 seconds for a typical head CT scan with 30 slices. Also, two-stage models are not significantly slower than single-stage models. This can be attributed to the highly optimized implementation of YOLOv5 which allows it to perform detection 5-10 times quicker (∼0.07 seconds) than the time taken for segmentation. We note that the high standard deviation throughout all prediction timings could be attributed to caching effects that occur at the start of the model inference. However, the mean prediction time reported remains representative of the average runtime for each slice.

IV. DISCUSSION

Overall, our results demonstrated the value of a two-stage approach where lesion detection is first performed before lesion segmentation. On our manually curated dataset of segmentation masks, the proposed two-stage approach was shown to outperform baseline models such as U-Net and state-of-the-art models such as SegResNet [31] by around 8%. We perceived that this performance improvement is due to the availability of larger datasets to train the first stage of the architecture. In the first stage, the model has learned

to propose the region of interest, alleviating the difficulty of training an end-to-end segmentation model on the whole image.

Ablation studies demonstrated that the introduction of the detection stage before segmentation gives a 4%-8% boost in segmentation performance as compared to direct segmentation, further affirming the value of our proposed two-stage approach. Interestingly, the YOLOv5-Seg model performed worse than SwinUnet and SegResNet. One of the reasons could be that the simple YOLOv5 segmentation head that up-samples the latent representations is less well-fitted than the other larger segmentation models on this task. Another key insight from the ablation study is the large increase in performance when using ground truth bounding boxes instead of those predicted by YOLOv5. There remains some room for improvement in terms of lesion detection performance (mAP@0.5 = 0.974, mAP@0.5:0.95 = 0.794) and the results obtained from using the ground truth show the best results that could be obtained by our proposed two-stage approach. If our two-stage model were to be implemented in clinical settings, neuroradiologists could make simple adjustments to the bounding box coordinates as a much quicker way of improving Dice scores (as compared to manually editing or creating the segmentation masks). These simple adjustments will also create improved bounding box labels that the model could use for fine-tuning the detection model and consequentially improve segmentation performance in the second stage, eventuating in the high Dice score of 0.769.

In terms of runtime, two-stage approaches that are supervised (i.e. YOLOv5 + TransDeepLab, YOLOv5 + SwinUnet) have insignificant overhead as compared to single-stage methods, making them equally suitable for clinical deployment. Even though a two-stage approach that uses unsupervised segmentation (i.e. YOLOv5 + EM) was able to outperform several single-stage segmentation models, the long runtime (approximately 10 times longer than many other non-EM approaches in Table 3) makes it less feasible for clinical deployment in emergency use cases. However, two-stage approaches with EM could potentially find applications in other less time-sensitive use cases where segmentation masks are not available at all.

Our proposed approach has been demonstrated for largely regular lesions and it might be less useful for segmenting lesions or structures that are large and irregular. For instance, segmentation masks of white matter or grey matter are spread across a large area and the resultant bounding box would have been large, limiting the utility of the first stage. On the other hand, future work could evaluate our two-stage approach on other datasets where lesions/structures are regular and segmentation masks are scant but bounding boxes can be easily demarcated, e.g. brain tumor segmentation, subcortical segmentation.

Another limitation of our approach is that our lesion dataset contains mostly one lesion subtype for each image. Therefore, we were able to obtain a bounding box in the ground truth mask that covers the lesion well. Future work in this research direction should robustly test more complicated scenarios such as CT images with multiple small lesions or multiple co-existing subtypes in the same slice.

Finally, our experiments are limited to 2D slices and future work in this area could explore the use of 3D datasets. This remains difficult considering the scarcity of datasets with segmentation masks. Alternatively, models that consider the sequential relationship across slices could be incorporated in the segmentation model in our two-stage framework, potentially improving the Dice scores further.

V. CONCLUSION

In this study, we have proposed a two-stage approach to ICH segmentation and demonstrated how it can leverage large datasets of bounding box annotations to improve downstream segmentation performance. This is especially useful considering how datasets of segmentation masks are typically very small. With this improved model, neuroradiologists could incorporate more accurate estimates of lesion measurements into their clinical decision-making process. Our proposed approach could also generalize to other medical image segmentation problems with small datasets.

REFERENCES

[1] J. J. Heit, M. Iv, and M. Wintermark, "Imaging of intracranial hemorrhage," *J. Stroke*, vol. 19, no. 1, pp. 11–27, Jan. 2017, doi: 10.5853/jos.2016.00563.

[2] O. Flower and M. Smith, "The acute management of intracerebral hemorrhage," *Current Opinion Crit. Care*, vol. 17, no. 2, pp. 106–114, Apr. 2011, doi: 10.1097/mcc.0b013e328342f823.

[3] K. Hu, K. Chen, X. He, Y. Zhang, Z. Chen, X. Li, and X. Gao, "Automatic segmentation of intracerebral hemorrhage in CT images using encoder–decoder convolutional neural network," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102352, doi: 10.1016/j.ipm.2020.102352.

[4] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018, doi: 10.1109/ACCESS.2017.2788044.

[5] M. Grewal, M. M. Srivastava, P. Kumar, and S. Varadarajan, "RADnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 281–284, doi: 10.1109/ISBI.2018.8363574.

[6] K. Kärkkäinen, S. Fazeli, and M. Sarrafzadeh, "Unsupervised acute intracranial hemorrhage segmentation with mixture models," in *Proc. IEEE 9th Int. Conf. Healthcare Informat. (ICHI)*, Aug. 2021, pp. 120–129, doi: 10.1109/ICHI52183.2021.00029.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, vol. 9351, 2015, pp. 234–241. [Online]. Available: https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[11] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis," *Intell. Med.*, vol. 3, no. 1, pp. 59–78, Feb. 2023, doi: 10.1016/j.imed.2022.07.002.

[12] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNET: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2023, pp. 205–218, doi: 10.1007/978-3-031-25066-8_9.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.

[14] R. Azad, M. Heidari, M. Shariatnia, E. K. Aghdam, S. Karimijafarbigloo, E. Adeli, and D. Merhof, "TransDeepLab: Convolution-free transformer-based DeepLab v3+ for medical image segmentation," in *Proc. 5th Int. Workshop Predictive Intell. Med.*, in Lecture Notes in Computer Science, Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, 2022, pp. 91–102, doi: 10.1007/978-3-031-16919-9_9.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[16] H. Ye, F. Gao, Y. Yin, D. Guo, P. Zhao, Y. Lu, X. Wang, J. Bai, K. Cao, Q. Song, H. Zhang, W. Chen, X. Guo, and J. Xia, "Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network," *Eur. Radiol.*, vol. 29, no. 11, pp. 6191–6201, Nov. 2019, doi: 10.1007/s00330-019-06163-2.

[17] W. Kuo, C. Häne, P. Mukherjee, J. Malik, and E. L. Yuh, "Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 45, pp. 22737–22745, Nov. 2019, doi: 10.1073/pnas.1908021116.

[18] D. Dong, G. Fu, J. Li, Y. Pei, and Y. Chen, "An unsupervised domain adaptation brain CT segmentation method across image modalities and diseases," *Expert Syst. Appl.*, vol. 207, Nov. 2022, Art. no. 118016, doi: 10.1016/j.eswa.2022.118016.

[19] X. Zhao, K. Chen, G. Wu, G. Zhang, X. Zhou, C. Lv, S. Wu, Y. Chen, G. Xie, and Z. Yao, "Deep learning shows good reliability for automatic segmentation and volume measurement of brain hemorrhage, intraventricular extension, and peripheral edema," *Eur. Radiol.*, vol. 31, no. 7, pp. 5012–5020, Jul. 2021, doi: 10.1007/s00330-020-07558-2.

[20] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[21] J. Yi, H. Tang, P. Wu, B. Liu, D. J. Hoeppner, D. N. Metaxas, L. Han, and W. Fan, "Object-guided instance segmentation for biological images," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12677–12684.

[22] A. Bhattacharya. (Feb. 20, 2024). *Brain Haemorrhage Segmentation From CT Scan Images Using Mask RCNN*. [Online]. Available: https://medium.com/datadriveninvestor/brain-haemorrhage-segmentation-from-ct-scan-images-using-mask-rcnn-e4f478ee10b2

[23] L. Jiang, D. Chen, Z. Cao, F. Wu, H. Zhu, and F. Zhu, "A two-stage deep learning architecture for radiographic staging of periodontal bone loss," *BMC Oral Health*, vol. 22, no. 1, p. 106, Dec. 2022, doi: 10.1186/s12903-022-02119-z.

[24] B. Bai, J. Tian, S. Luo, T. Wang, and S. Lyu, "YUSEG: YOLO and UNet is all you need for cell instance segmentation," in *Proc. Cell Segmentation Challenge Multi-Modality High-Resolution Microsc. Images*, 2023, pp. 1–15.

[25] Y. Baba and A. Murphy, (2017), "Windowing (CT)," *Radiopaedia.org*, doi: 10.53347/rID-52108.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[27] G. Jocher, A. Stoken, and J. Borovec. (Feb. 11, 2022). *Ultralytics/YOLOv5: V5.0—YOLOv5-P6 1280 Models, AWS, Supervisely and YouTube Integrations*. [Online]. Available: https://github.com/ultralytics/yolov5/tree/v5.0

[28] G. Jocher et al., (2021), "Ultralytics/YOLOv5: V5.0—YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations," *Zenodo*, 2021, doi: 10.5281/zenodo.4679653.

[29] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier, "Development and validation of deep learning algorithms for detection of critical findings in head CT scans," in *Proc. Comput. Vis. Pattern Recognit.*, Mar. 2018.

[30] A. E. Flanders et al., "Construction of a machine learning dataset through collaboration: The RSNA 2019 brain CT hemorrhage challenge," *Radiol., Artif. Intell.*, vol. 2, no. 3, May 2020, Art. no. e190211, doi: 10.1148/ryai.2020190211.

[31] E. P. Reis, F. Nascimento, M. Aranha, F. M. Secol, B. Machado, M. Felix, A. Stein, and E. Amaro, (2020), "Brain hemorrhage extended (BHX): Bounding box extrapolation from thick to thin slice CT images (version 1.1)," *PhysioNet*, doi: 10.13026/9cft-hg92.

[32] (Jan. 22, 2023). *SingHealth*. [Online]. Available: https://www.nni.com.sg/

[33] R.-G. Dumitru, D. Peteleaza, and C. Craciun, "Using DUCK-Net for polyp image segmentation," *Sci. Rep.*, vol. 13, no. 1, p. 9803, Jun. 2023, doi: 10.1038/s41598-023-36940-5.

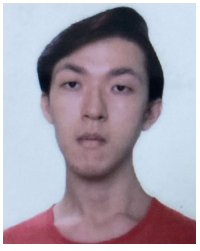
[34] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *Proc. 4th Int. Workshop*, 2019, pp. 311–320, doi: 10.1007/978-3-030-11726-9_28.

[35] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[36] M. F. Sharrock, W. A. Mould, H. Ali, M. Hildreth, I. A. Awad, D. F. Hanley, and J. Muschelli, "3D deep neural network segmentation of intracerebral hemorrhage: Development and validation for clinical trials," *Neuroinformatics*, vol. 19, no. 3, pp. 403–415, Jul. 2021, doi: 10.1007/s12021-020-09493-5.

JAGATH C. RAJAPAKSE (Fellow, IEEE) was a Visiting Professor with Massachusetts Institute of Technology (MIT), USA, a Visiting Scientist with the Max Planck Institute for Human Cognitive and Brain Sciences, Germany, and the National Institutes of Health, USA. He is currently a Professor of computer engineering with NTU, Singapore. His research interests include deep learning, brain imaging, and biological networks. He has published more than 300 peer-reviewed articles in these areas. He has served as an Associate Editor for IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS.

CHUN HUNG HOW received the Bachelor of Science degree in mathematical sciences and economics from NTU, Singapore, where he is currently pursuing the M.Eng. degree in computer science. He works in the area of computer vision applied to medical imaging data. His work mainly focuses on student-teacher learning, generative models, and multitask segmentation models.



YI HAO CHAN received the Bachelor of Engineering degree in computer science from NTU, Singapore, where he is currently pursuing the Ph.D. degree in computer science. He works in the area of medical image analysis, including structural, diffusion, and functional MRI. His work has mainly focused on developing explainable AI models to discover multimodal disease biomarkers from morphological, structural, and functional connectomes.

LUKE CHIN PENG HAO received the Bachelor of Engineering degree in computer science from NTU, Singapore. He is currently becoming a Software Engineer in the private sector. During the bachelor's degree, he worked on projects relating to NLP and image recognition.

ABHINANDAN PADHI received the Bachelor of Engineering degree in computer science from NTU, Singapore. Since 2022, he has been a Data Scientist in a Singapore-based AI startup. He has coauthored a paper in the 2021 International Conference on Data Mining and the Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE) Workshop Series. His research interests include artificial intelligence and data science, specifically deep learning applications in computer vision and natural language processing.

VIVEK ADRAKATTI received the Bachelor of Engineering degree in computer engineering from NTU, Singapore. He is currently pursuing the M.Sc. degree in computer science with the University of Illinois at Urbana–Champaign, Champaign, IL, USA. He has completed past internships with Amazon, Panasonic Research and Development Center, Shopee, and IBM, as a Software Developer, a Data Scientist, and a Machine Learning Engineer. He is from and completed his schooling in Bengaluru, India. His research interests include the application of deep learning to medical imaging and human–computer interaction. He has been awarded the NTU President's Research Scholar Award two times during the bachelor's degree.

IRAM RAIS ALAM KHAN received the M.B.B.S. degree from the MGM Medical College, Navi Mumbai, in 2010, the D.N.B. degree in radiodiagnosis from the Breach Candy Hospital and Research Centre, Mumbai, in 2014, and the M.Med. degree in diagnostic radiology from the National University of Singapore, in 2017. She was a Senior Staff Registrar with the Department of Neuroradiology, National Neuroscience Institute, Singapore, in 2022. She is currently a fellow with the Royal College of Radiologists (FRCR), U.K. Her research projects and interests include head and neck imaging and neuro-oncology imaging. In 2016, she achieved a Fellowship with Ultrasound and Color Doppler from the T. N. Medical College and the B. Y. L. Nair Charitable Hospital, Mumbai. She received the Fellowship in Neuroradiology from the National Neuroscience Institute, in 2017. She has been a Reviewer of manuscripts in the journal *Insights into Imaging*.

TCHOYOSON LIM received the M.B.B.S. and Master of Medicine degrees from the National University of Singapore, Singapore. He was an Adjunct Associate Professor of diagnostic radiology with the Yong Loo Lin School of Medicine, National University of Singapore. Since 2005, he has been a Senior Consultant with the National Neuroscience Institute, Singapore. He is currently an Adjunct Associate Professor with the Duke-NUS Graduate Medical School. He is also a fellow with the Royal College of Radiologists, U.K., and the Academy of Medicine, Singapore. He has coauthored more than 100 peer-reviewed articles and holds two U.S. patents in medical image archiving. He served on the Editorial Board for many journals, including *The British Journal of Radiology*, *Annals of the Academy of Medicine*, and *Singapore Medical Journal*. He has been a Reviewer of numerous journals, such as *The British Journal of Radiology*, *European Radiology*, *Neuroradiology*, *Magnetic Resonance Imaging*, and *Journal of Neurology, Neurosurgery, and Psychiatry*. He served as the Secretary for Singapore Radiological Society.