Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2024, Article ID 3819801, 14 pages
https://doi.org/10.1155/2024/3819801

Research Article

YOLO-UNet Architecture for Detecting and Segmenting the Localized MRI Brain Tumor Image

Nur Iriawan,[1] Anindya A. Pravitasari,[2] Ulfa S. Nuraini,[1,3] Nur I. Nirmalasari,[1] Taufik Azmi,[1] Muhammad Nasrudin,[1] Adam F. Fandisyah,[1] Kartika Fithriasari,[1] Santi W. Purnami,[1] Irhamah,[1] and Widiana Ferriastuti[4]

[1]*Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, East Java, Indonesia*
[2]*Department of Statistics, Universitas Padjadjaran, Bandung, West Java, Indonesia*
[3]*Data Science Department, Universitas Negeri Surabaya, Surabaya, East Java, Indonesia*
[4]*Department of Radiology, Universitas Airlangga, Surabaya, East Java, Indonesia*

Correspondence should be addressed to Anindya A. Pravitasari; anindya.apriliyanti@unpad.ac.id

Received 19 December 2022; Revised 13 August 2023; Accepted 25 January 2024; Published 8 February 2024

Academic Editor: Ridha Ejbali

Brain tumor detection and segmentation are the main issues in biomedical engineering research fields, and it is always challenging due to its heterogeneous shape and location in MRI. The quality of the MR images also plays an important role in providing a clear sight of the shape and boundary of the tumor. The clear shape and boundary of the tumor will increase the probability of safe medical surgery. Analysis of this different scope of image types requires refined computerized quantification and visualization tools. This paper employed deep learning to detect and segment brain tumor MRI images by combining the convolutional neural network (CNN) and fully convolutional network (FCN) methodology in serial. The fundamental finding is to detect and localize the tumor area with YOLO-CNN and segment it with the FCN-UNet architecture. This analysis provided automatic detection and segmentation as well as the location of the tumor. The segmentation using the UNet is run under four scenarios, and the best one is chosen by the minimum loss and maximum accuracy value. In this research, we used 277 images for training, 69 images for validation, and 14 images for testing. The validation is carried out by comparing the segmentation results with the medical ground truth to provide the correct classification ratio (CCR). This study succeeded in the detection of brain tumors and provided a clear area of the brain tumor with a high CCR of about 97%.

1. Introduction

A brain tumor is the 15th deadly disease with a high mortality rate in Indonesia in 2018. According to the Global Cancer Observatory [1], there were 4,229 mortalities in 5,323 cases of brain and nervous system tumors. The high accuracy of medical treatment is needed since the brain tumor and healthy part are not clearly separated; therefore, the clear shape, boundary, and location of the brain tumor are useful information to increase the safety probability, especially in medical surgery. The challenge in providing that information is mostly caused by the heterogeneous appearance of the tumor and the quality of MRI images. Manual detection and segmentation are time-consuming due to a large number of MRI images and their error-prone to human subjectivity. Therefore, a more objective and effective approach is needed to deal with this issue.

The detection of brain tumor images through classification and segmentation has been widely proposed by the supervised and unsupervised learning approaches. Unsupervised learning is powerful due to the data number limitation. However, unsupervised learning does not take any feedback to check if the prediction provides the correct output. It is more about finding the hidden patterns in the data. In the current state, supervised learning is favorable since the training section considered provides better results. Automatic detection with deep learning is a hot issue and has shown ground-breaking performance in a variety of sophisticated image tasks. Various methodologies and architectures of deep learning are developed in the image

processing area, such as CNN and FCN. CNN is powerful and widely used in classification, while FCN is greatly used in semantic segmentation.

CNN applications in medical image processing have been extensively investigated across diverse contexts and architectural frameworks. One study focused on brain tumor classification implemented a deep CNN with transfer learning [2]. Another research endeavor employed a novel architecture termed multikernel depthwise convolution for chest X-ray image analysis [3]. The classification of brain diseases was addressed using a landmark-based deep learning approach with the VGG-f architecture as a pre-trained model [4]. Deep ensemble learning on MRI brain images was explored through the DeepESRNet architecture [5]. Additionally, various CNN architectures, including AlexNet, VGG-19, and ResNet, were utilized to categorize osteoarthritis in the hip joint based on X-ray images [6]. Object detection in both real-time and artwork images was achieved using the YOLO model [7], which was then compared with other object detectors such as R-CNN, faster R-CNN, and the deformable parts model (DPM). All previous architecture has great results in classification; however, the time processing issue is also important besides its high accuracy. This study uses the CNN-YOLO (You Look Only Once) model since it could be dealing with this issue. In classifying, the YOLO could localize objects in the input image and process the entire image at once, unlike the sliding window approach that is followed by the convolutional neural network (CNN) architecture [8]. This is why the YOLO model proved to be much faster.

This study has the purpose of providing the automatic detection of a brain tumor and then segmenting it to produce a clear shape and boundary of the tumor. The UNet architecture of FCN is used for semantic segmentation [9–13] since it has great performance on very different biomedical segmentation applications and power powerful for limited training number images. However, its complex architecture makes UNet less desirable. Therefore, this study generates four scenarios from the UNet to overcome its complexity and choose the best scenarios with the loss and accuracy criteria. Another study that employed automatic detection [14] used the R-CNN in brain tumor detection and segmentation. The mask R-CNN was introduced to detect multiple sclerosis lesions [15, 16]. The comparisons of several studies are summarized in Table 1.

The automatic detection and clear segmentation results are the main purposes of this research. An MRI brain image of a patient is a group of images (of about 25 to 30 image slices) assembled from scans of the head beginning at the pharynx and ending at the tip of the skull. In conventional segmentation, doctors and paramedics must manually select which slice contains the tumor, which is time-consuming. The serial combination of YOLO and UNet, namely YOLO-UNet, is the way for resolving manual detection in segmenting the serial MRI image. Moreover, this combination has main advantages which could eliminate the preprocessing step, because the YOLO could localize the tumor area in the MRI image that impacts the UNet segmentation process more focused on the tumor area and avoids noise that interferes with segmentation. This kind of analysis provided automatic detection and segmentation as well as the location of the tumor. In addition, the results of the segmentation of the proposed model are compared with the previous method, i.e., the UNet and mask R-CNN, which have the same implementation to classify individual objects and localize them using bounding boxes and semantic segmentation, respectively. The correct classification ratio (CCR) will be calculated to compare the segmentation results with the medical ground truth as the measure of evaluation. The training and validation datasets are provided online while the testing dataset is from General Hospital Dr. Soetomo Surabaya, and it is restricted due to the medical ethical clearance privacy policy.

Managing the various levels and types of noise present in real-world MRI images is another significant challenge in the field of medical imaging. MRI images are essential for diagnosing and monitoring various medical conditions but can be affected by various noise sources that can degrade image quality and impact diagnostic accuracy. A limitation of this study is the scarcity of datasets that encompass various types of noise present in real-world MRI images. Nevertheless, the research strives to tackle these challenges by generating synthesized noise, in order to thoroughly evaluate the given approach.

2. Proposed YOLO-UNet for Automatic Detection Segmentation

The YOLO-UNet model for automatic detection segmentation is the serial combination of YOLO and UNet architecture. YOLO (You Only Look Once) is an approach in deep learning that performs object detection. It is targeted at real-time processing and framing objects as a single regression problem from direct image pixels to separate spatial bounding boxes and associated probability classes. YOLO performs object detection and recognition like the human brain. When humans look at something, the brain instantly recognizes and makes a conclusion about what is being seen. In detecting an object, the classifiers that are utilized by the current detection framework take the classifier to a particular object and evaluate it on various scales and locations in the testing image [17]. This study used YOLO architecture since it is very fast and accurate. YOLOv3 [18] and YOLOv4 [19] are the common variants of YOLO based on the Darknet architecture that are easy to combine with UNet and more compatible with our computer specifications. To assess the effectiveness of both architectures in detecting tumors, we will compare YOLOv3 and YOLOv4. The YOLO itself will be combined with the several scenarios of the UNet architecture. The combinations were then compared to each other to investigate the best model.

YOLOv3 has a total of 53 convolutional layers; therefore, this architecture is also known as Darknet-53 [18]. In YOLOv3, a convolutional layer is always followed by batch normalization and leaky ReLU. Residual block or shortcut connection on YOLOv3 is carried out by adding up the inputs before the convolutional layer residual block with the results from the 1×1 convolutional layer filter followed by batch normalization and leaky ReLU, followed by a 3×3

Table 1: Comparison of several studies for classification and segmentation in medical image processing.

Author	Methods	C*	S*	ADS*	Dataset	Size	Type
Deepak and Ameer [2]	GoogLeNet with transfer learning	√	—	—	3064 brain MRI images from 233 patients (diagnosed meningioma, glioma, and pituitary tumors) are with T1-CE MRI modality and include coronal, sagittal, and axial views	512 × 512	.Mat file
Hu et al. [3]	Multikernel depthwise convolution	√	—	—	Chest X-ray 14 dataset and Chest X-ray 2017 dataset	512 × 512 preprocess to 256 × 256	Dicom
Liu et al. [4]	Landmark-based deep multi-instance learning (LDMIL)	√	—	—	1.5T and 3T T1-weighted structural MRI data (1526)	Voxel level with 256 × 256 × 256	—
Suk et al. [5]	Deep ensemble learning of sparse regression models	√	—	—	805 subjects of 186 (AD), 393 (MCI), and 226 (normal control, NC)	Voxel level with 256 × 256 × 256	Neuroimaging Informatics Technology Initiative (NIfTI)
Xue et al. [6]	AlexNet, VGG-19, and ResNet	√	—	—	420 anteroposterior (AP) view hip X-ray images with normal (219) and OA (201)	432 mm × 356 mm, with a pixel resolution of 0.187 mm × 0.187 mm	—
Safdar et al. [8]	YOLO	√	—	—	T1-weighted and FLAIR images from MRI images of patients who suffered from low-grade glioma	256 × 256	Dicom
Ronneberger et al. [9]	UNet	—	√	—	Transmission electron microscopy of the Drosophila first instar larva ventral nerve cord (VNC)	512 × 512	—
Weng et al. [11]	UNet	—	√	—	MRI, CT, and ultrasound (Promise12, chaos, NERVE dataset)	256 × 256; 512 × 512; 580 × 420	Dicom
Li et al. [12]	UNet	—	√	—	NIH pancreatic segmentation dataset that contains 82 contrast-enhanced abdominal CT volumes and corresponding fine annotations	512 × 512 × L, where L ∈ [181, 466]	—
Pravitasari et al. [10]	UNet-VGG16	—	√	—	Real dataset from the general hospital (RSUD) Dr. Soetomo (152)	256 × 256	Dicom
Bisa [14]	R-CNN	√	√	√	BraTS (brain tumor segmentation) challenge dataset	—	Neuroimaging Informatics Technology Initiative (NIfTI)
Süleyman Yıldırım and Dandıl [15]	Mask R-CNN	√	√	√	MR image of multiple sclerosis dataset (1838)	min 256, max 512	—
He et al. [16]	Mask R-CNN	√	√	√	COCO dataset	640 × 800	JPG

*Note: C = classification, S = segmentation, and ADS = automatic detection segmentation.

convolutional layer filter with batch normalization, and at the end, a leaky ReLU is performed. YOLOv4 is the improved version of YOLOv3 which has great performance in speed and accuracy. The features added in YOLOv4 are two methods called Bag of Freebies (BOF) and Bag of Special (BOS). Both of them are applied to the detector module's backbone. The detection heads of YOLOv4 and YOLOv3 are similar, but it has three feature maps at different levels of the convolutional procedure. This makes YOLOv4 have a total of 161 network layers that can improve the accuracy compared to YOLOv3.

The YOLO is used to perform the classification and localization of the tumor coordinates. The illustration of the classification procedure for localizing the brain tumor image is shown briefly in Figure 1.

3. Materials and Methods

3.1. Dataset. The dataset used in this study contains 346 images of axial MRI slices. The 277 images for the training and validation sections are provided online by Bhuvaji et al. [20] from Kaggle (https://www.kaggle.com/sartajbhuvaji/brain-tumor-classification-mri), while the rest of images for the testing section are from Dr. Soetomo. The dataset from Bhuvaji et al. provides various types of tumors, including glioma tumors, meningiomas, and pituitary tumors. The tumor images provided by this dataset exhibit a wide range of variations in terms of size and tumor location. This study does not classify the tumor images; rather, it focuses on tumor detection and segmentation. Consequently, the challenge at hand is to design a proposed model capable of detecting tumor regions regardless of their size or location.

The dataset contains MRI grayscale images with various sequence names T1, memp + C, and T2 FLAIR. The T1 memp + C sequence is the slices that have been added with a contrast media. This made the tumor segment more visible. The T2 FLAIR, on the other hand, is a sequence without contrast media, in which the more visible feature is the swelling or edema. The training dataset is only chosen from the axial point of view. This is carried out to match the testing data from Dr. Soetomo with only available in axial. The summary of the dataset is given in Table 2.

The difference between data training, validation, and testing must be equated to facilitate the analysis. Therefore, the testing data is converted to *.jpg type, and the dimension is lumped to 256×256. Preprocessing is carried out initially to normalize the variables before the automated detection and segmentation. In the training process, the dataset did not take the augmentation process since we wanted to train the architecture with the real dataset. This simple preprocessing is taken for a reason to prove that the localized image by YOLO could substitute the preprocessing step that is commonly used. However, to add more challenge in the testing process, we did some augmentation by generating synthetic noise, i.e., Gaussian and speckle noise.

The dataset ground truth is created by medical judgment from the radiologist. The training and validation sections were carried out in serial by YOLO and UNet. The output localizes the image that contains the brain tumor and then considers as the training data for the UNet. YOLOv3 and YOLOv4 are the models used in this study; moreover, the UNet is run under four scenarios. The best-pretrained models are chosen from several combinations of YOLO and UNet models based on loss and accuracy metrics. The testing data will use the best pretrained model of YOLOv3-UNet and YOLOv4-UNet. All of the models were created using the Tensorflow, Keras, and NumPy libraries and the Python programming language, which is run on a computer with an Intel Core i7 CPU, 32 GB of RAM, a 128 GB SSD, and no GPU or VRAM. As an experimental result, the proposed model also compares with another method, i.e., UNet and mask R-CNN. The use-case diagram for the analysis is given in Figure 2.

3.2. Hyperparameter Setting. Based on Figure 1, the image input has a dimension of 256×256. The pixels that have been rescaled are in accordance with the hyperparameter, to obtain the classification and localization of brain tumors on a scale of 13×13, 26×26, and 52×52. The YOLOv3 and YOLOv4 hyperparameter settings for this study are shown in Table 3.

Localize brain tumor area as the output of YOLO becomes the input of UNet architecture. UNet is one of the FCN architectures for image segmentation [9]. Its goal is to predict each pixel's class. The UNet network architecture consists of a down-sampling (encoding) path and an up-sampling (decoding) path. The down-sampling path has 6 convolutional blocks. Each convolutional block has 2 layers, and every layer has filters. The number of feature maps is increasing from the original size. In every up-sampling block, two convolutional layers reduce the number of feature maps. In down-sampling and up-sampling, the path used "same" padding for all convolution layers. This complex structure of UNet has a disadvantage for execution, especially in running time. Moreover, the complex architecture is not always compatible with all computer specifications.

This study will use four UNet architectural scenarios and compare them with loss and accuracy measurements. The number of convolution layers and convolution blocks used is very influential in finding the best model. The number of parameters and feature maps that are formed are also affected by the convolution layer and block. Table 4 shows a breakdown of the number of convolution layers and convolution blocks in each scenario, and the model architecture differences can be seen in Figure 3.

Figure 3(a) shows architectural visualization for model 1, and Figures 3(b)–3(d), respectively, show architectural visualizations for models 2, 3, and 4. Model 1 and model 2 have one convolution layer for each convolution block. This causes a reduction in the number of parameters more than half of the model than that had by the two convolution layers for each convolution block, namely, model 3 and model 4. In model 1 and model 3, five convolution blocks are used at the encoder stage. Each output of the convolution block stage will produce a matrix with a size half of the original size, a 1/32 matrix is

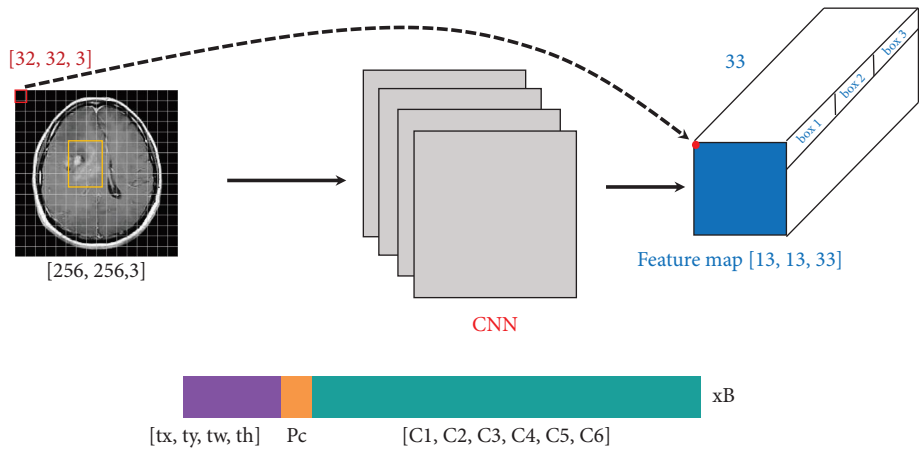


Figure 1: Classification and localization of the tumor coordinates.

Table 2: Summary of the dataset.

Source	Number of data	Type	Dimension (pixel)	Used for
Online	277	*.jpg	256 × 256	Training
Online	69	*.jpg	256 × 256	Validation
Dr. Soetomo Hospital	14	*.dicom	256 × 256 and 512 × 512	Testing

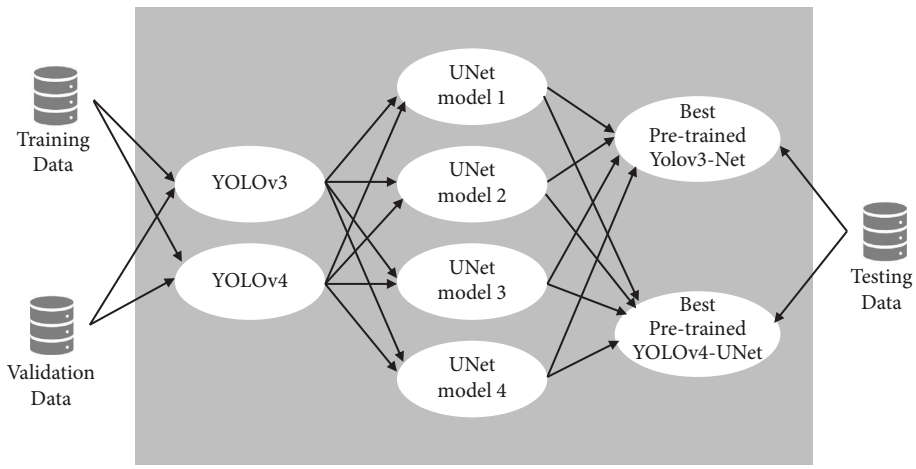


Figure 2: The use-case diagram.

Table 3: Hyperparameter setting.

Model	Hyperparameter	YOLOv3	YOLOv4
Darknet-53	Image size	256 × 256	256 × 256
	Batch size	64	64
	Subdivisions	16	16
	Training step	6000	6000
	Learning rate	0.001	0.001
Model	Anchors	(10 × 13)	(12 × 16)
		(16 × 30)	(19 × 36)
		(33, 23)	(40, 28)
		(30, 61)	(36, 75)
		(62, 45)	(76, 55)
		(59, 119)	(72, 146)
		(116, 90)	(142, 110)
		(156, 198)	(192, 243)
		(373, 326)	(459, 401)

Table 4: Convolution scenario.

Scenario	Number of data		
	Convolutional layer	Convolutional block	Parameter
Model 1	1 conv @ block	10 block	3,930,273
Model 2	1 conv @ block	8 block	980,385
Model 3	2 conv @ block	10 block	7,862,401
Model 4	2 conv @ block	8 block	1,962,625

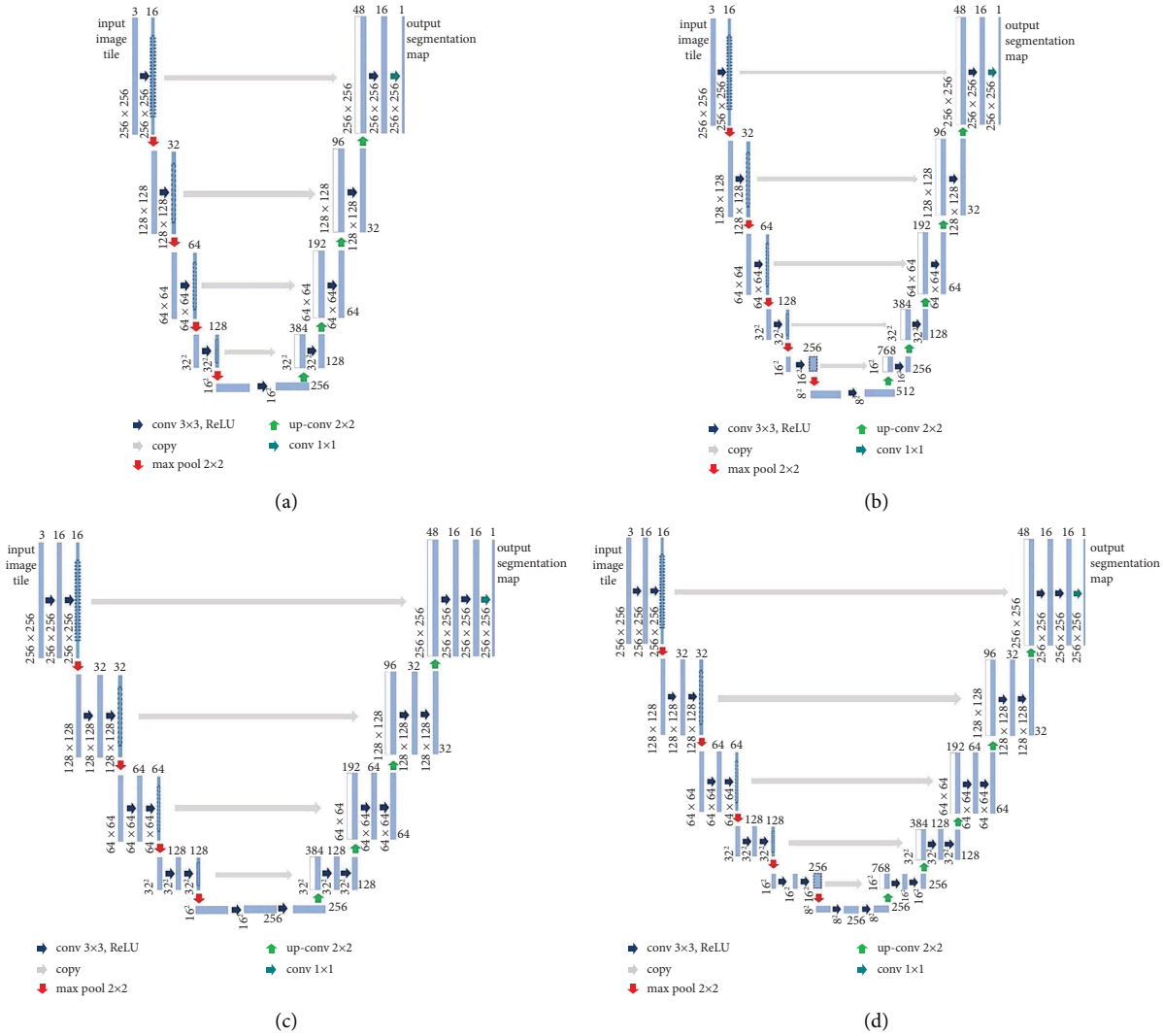


Figure 3: The architectural visualization of each scenario: (a) model 1, (b) model 2, (c) model 3, and (d) model 4.

obtained at the end of the encoder stage (8×8 dimensions of a matrix). Model 2 and model 4 have four convolution blocks at the encoder stage which cause the final matrix to be 1/16 of the original size. The final matrix for model 2 and model 4, therefore, will be at 16×16 dimensions.

3.3. Model Evaluation. The best model is chosen based on several metrics' evaluations. The first is the YOLO section, which uses the mean average precision (mAP) as the evaluation of the precision of detection results. The mAP value is the average value of average precision. Each precision value is calculated from each item generated by the system after being sorted. It is simply calculated with the following formula [21]:

$$\mathrm{mAP} = \frac{1}{c} \sum_{i=1}^{c} \mathrm{AP}_i. \tag{1}$$

The AP is calculated with the all-point interpolation method as follows:

$$\mathrm{AP} = \sum_{n=0} (r_{n+1} - r_n) p_{\mathrm{interp}}(r_{n+1}),$$
$$p_{\mathrm{interp}}(r_{n+1}) = \max_{r:r \geq r_n+1} p(\tilde{r}), \tag{2}$$

where C is the number of classes, p is precision, p_{interp} is precision interpolation, r is recall, and $p(\tilde{r})$ is the precision calculated by the recall. AP value can be obtained by providing the intersection over union (IoU) value. The IoU measures the overlap between 2 boundaries. The IoU will be used to determine whether the predicted bounding box is true positive (TP), false positive (FP), or false negative (FN).

The second section is the YOLO-UNet comparison. This stage uses the loss and accuracy from UNet performance criteria. The loss represents the value of error between predicted and actual. The case with two classes in machine learning uses the binary cross-entropy loss function to calculate the value of loss or error [22]. The binary cross-entropy is provided by the following equation:

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^{N} z_i \log(p(z_i)) + (1 - z_i) \log(1 - p(z_i)), \tag{3}$$

where N is the number of data, z_i is the class of classification which has the value of 0 or 1, and $p(z_i)$ is the probability of z_i. The formula for calculating the accuracy is shown by the following equation:

$$\mathrm{accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}. \tag{4}$$

The last section is the experimental results by comparing the segmentation results from testing data with another state-of-the-art architecture. This study uses the correct classification ratio (CCR) to determine whether the region of interest (ROI) from the segmentation results is in accordance with the ground truth. The greater the CCR value, the better the segmentation result [23, 24]. The CCR can be calculated by the following equation:

$$\mathrm{CCR} = \sum_{i=1}^{2} \frac{\left|\mathrm{GT}_j \cap \mathrm{Seg}_j\right|}{|\mathrm{GT}|}, \tag{5}$$

where GT_j is ground truth, Seg_j is the segmented pixel, and $\mathrm{GT} = \bigcup_{j=1}^{2} \mathrm{GT}_j$. The index $j = 1$ and $j = 1$ denote the non-ROI area and ROI area, respectively.

The experimental setup in this study can be visualized in Figure 4. In the activity diagram, it can be seen that detection and segmentation are performed serially by applying 2 YOLO scenarios to detect tumors and segmenting them with 4 UNet scenarios. The combination of YOLO-UNet produces 8 models that are validated based on evaluation metrics and CCR.

4. Results and Discussion

4.1. Training Section. The training sections are carried out separately for YOLOv3 and UNet. The YOLOv3 runs under the condition of a hyperparameter as shown in Table 3. For the training dataset, both YOLOv3 and YOLOv4 provide an mAP perfect score of about 100% as shown in Figure 5. As a result, both architectures are taken into consideration when combining the four UNet scenarios.

A 100% mAP (mean average precision) score signifies the precision of the model's prediction in terms of bounding boxes and class probabilities in the ground truth annotations across all images in the training dataset. This comparison is based on the IoU (intersection over union) between predicted boxes and ground truth boxes, resulting in the values of TP (true positives—correctly predicted bounding boxes), FP (false positives—predicted bounding boxes that are incorrectly predicted), and FN (false negatives—predicted bounding boxes that are incorrectly missed). Therefore, achieving a 100% accuracy indicates that the employed model has accurately and effectively targeted the detection and classification of tumors.

The localized brain tumor images become the training input for UNet. Training with the UNet model runs under four scenarios as shown in Table 4 and Figure 4. All models use the "same" padding in the encoder and decoder parts for all convolution layers. Backpropagation operations are performed for each update of each epoch for the calculation of the accuracy and loss of training data and validation. It improves the value of loss and accuracy in the model for each epoch. Figure 6 demonstrates graphs of the loss value and accuracy of the validation data for each model under YOLOv3 and YOLOv4.

From Figure 6, it is known that the loss value of each model decreases close to 0, and the accuracy of each model increases up to 0.99. The minimum the loss value produced, the better the classification results are obtained. The loss and accuracy of YOLOv3-UNet are better than YOLOv4-UNet. It can be seen with a small loss and a higher accuracy. Table 5 shows the result of loss and accuracy in the training data and validation at the last epoch.

Table 5 shows the minimum loss and maximum accuracy for YOLOv3-UNet in training and validation data. The best model in training data is model 2, while the best model in the validation data is provided by model 3. YOLOv4-UNet has different results. For training data, the best model is reached by model 4. The various results are provided by the validation data, which show the minimum loss in model 2 and the maximum accuracy in models 1, 3, and 4. Based on these results, the best models were determined by the minimum loss for all models; therefore, model 3 in YOLOv3-UNet and model 2 in YOLOv4-UNet were chosen. Both models will be used for the next analysis.

4.2. Testing Section. The testing section used 14 datasets from Dr. Soetomo. To add more challenge to the experimental results, the 14 original datasets were added by Gaussian and speckle noise. All datasets are modeling with YOLOv3-UNet and YOLOv4-UNet compared with UNet (four model scenarios in Table 4, without YOLO) and mask R-CNN. Figure 7 shows some results of segmentation with all models.

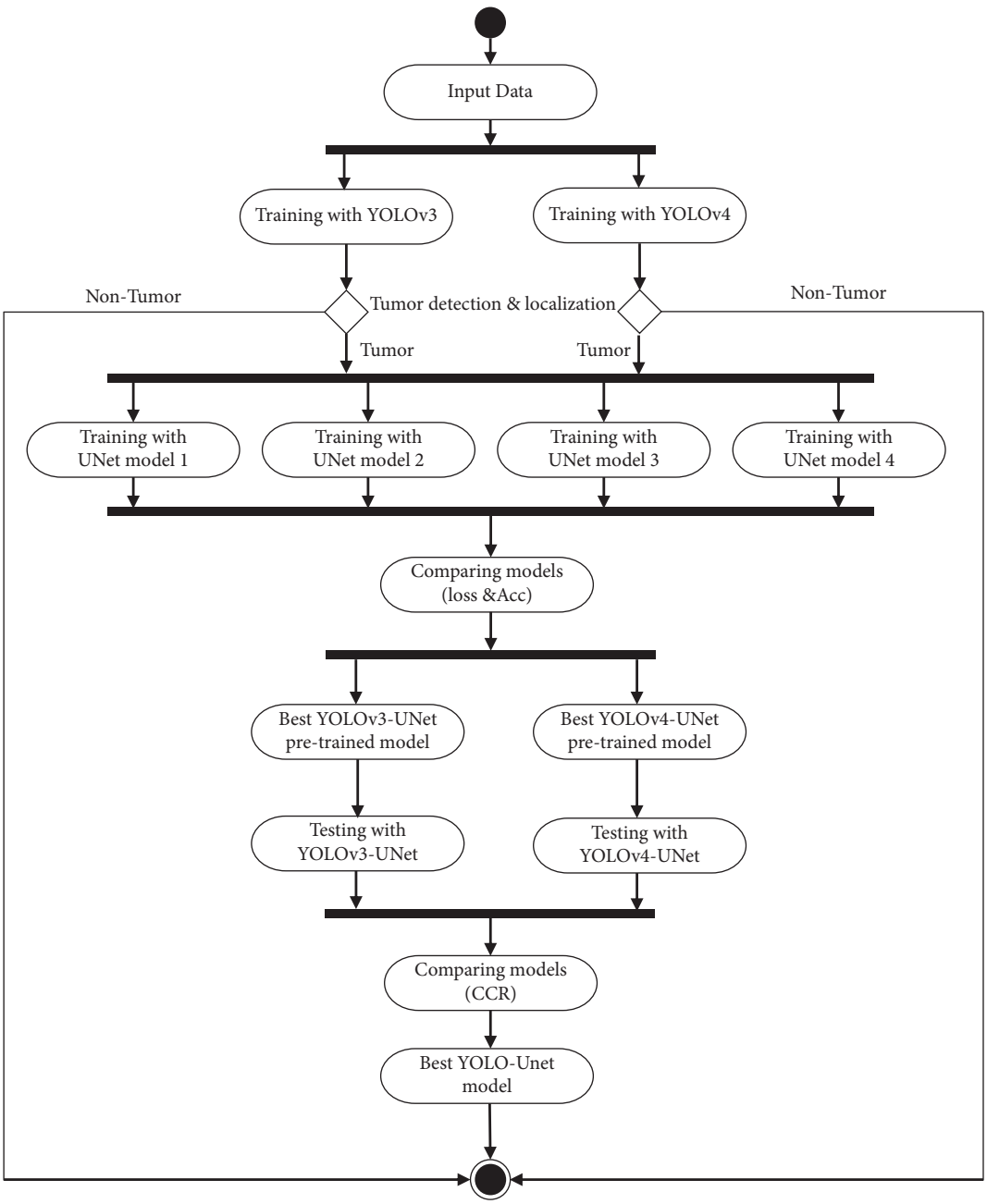
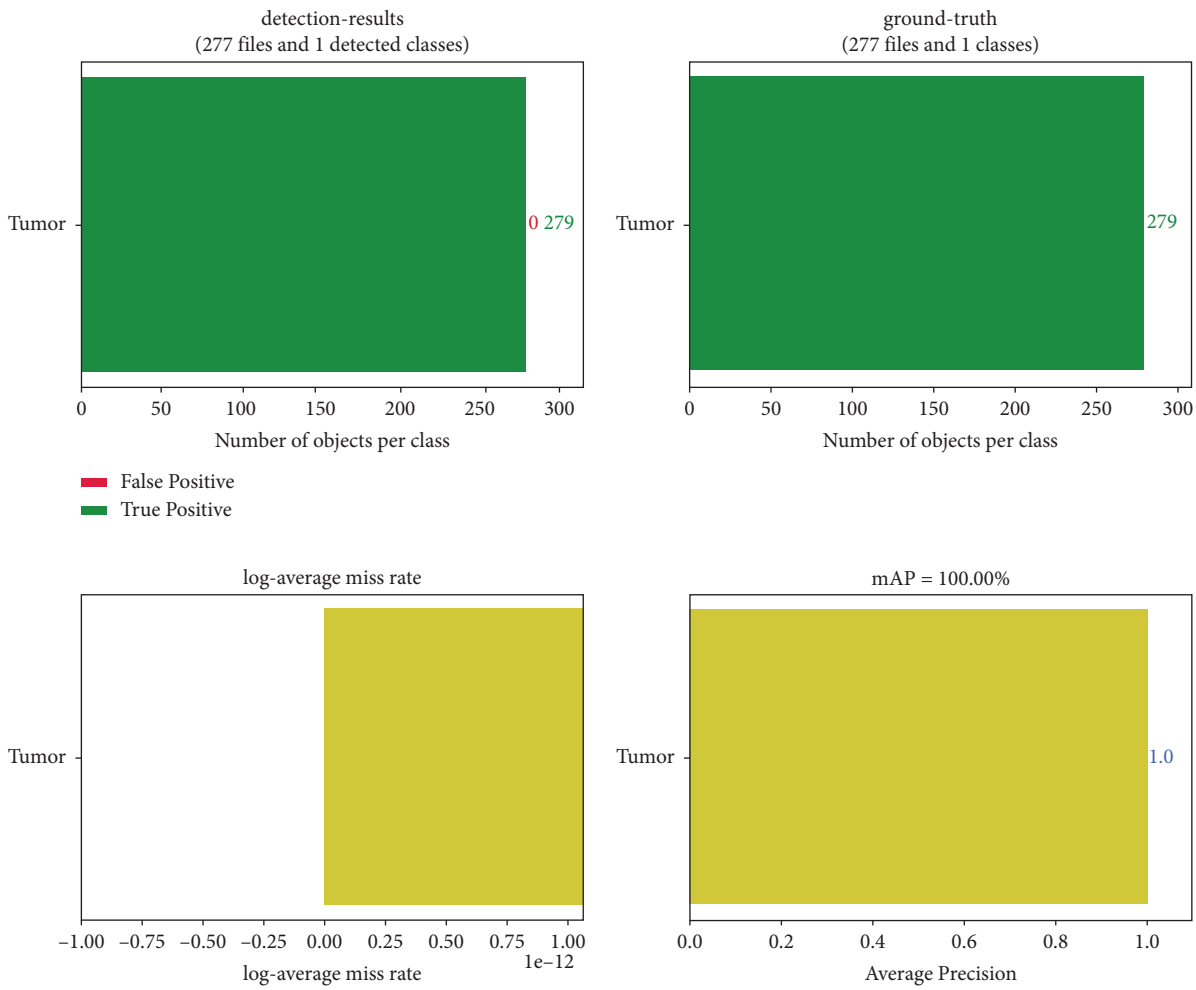


Figure 4: The activity diagram.

The numerical comparison of each model is shown in Table 6. It is the calculation of the average CCR for each model and each dataset. From Table 6, it is clear that YOLOv3-UNet is great in segmenting the original image and the Gaussian noising dataset. However, for the dataset with speckle noise, UNet model 3 has the best accuracy among the other models. Yolov4-UNet gave unsatisfaction results, and this model could not even detect the tumor area in a dataset with Gaussian noise. The mask R-CNN could provide great segmentation results with a consistent CCR of about 96% per dataset.

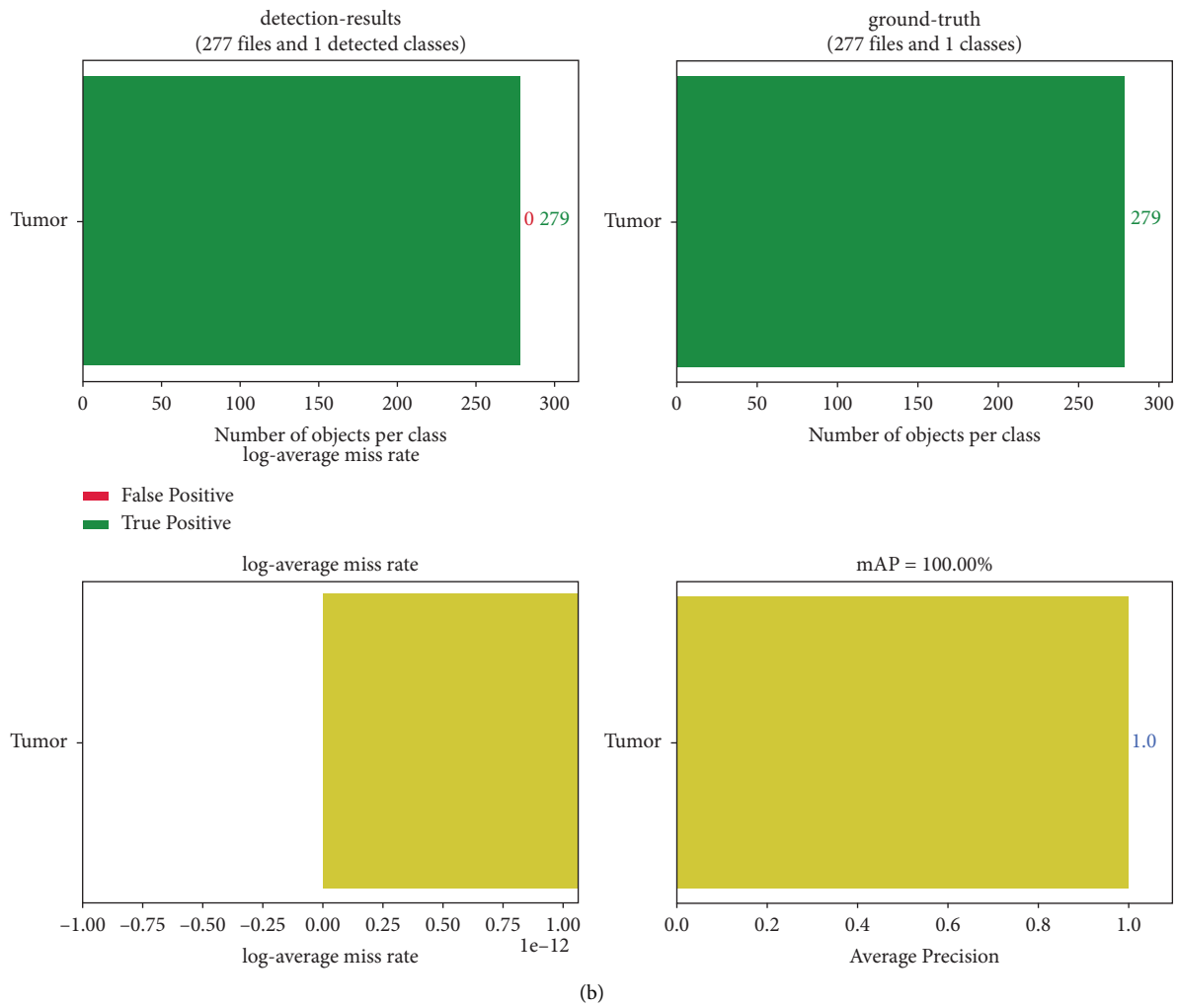
5. Discussion

This study's main contribution is combining the CNN-FCN methodology, namely, YOLO-UNet architecture, to provide automatic detection and segmentation. Several studies have been conducted through detection followed by classification which is carried out separately from segmentation. This study combines two kinds of analysis and could provide significant results. Yolov3 and YOLOv4 are chosen since it has great performance and work very fast than the others. The UNet architecture



(a)

Figure 5: Continued.



(b)

Figure 5: The precision of training dataset: (a) YOLOv3 and (b) YOLOv4.

complexity is overcome by creating four scenarios, and the best model is chosen by the maximum value of accuracy and minimum value of loss.

Several challenges in terms of the dataset arise when performing tumor detection. The varying sizes and locations of different tumor types necessitate the developed model to recognize and detect these tumors accurately. Several models were experimented with, including YOLOv3 and YOLOv4, for this purpose. Due to computational limitations, these two models were ultimately chosen for implementation. In the context of detection, this approach is considered capable of addressing various types of brain tumors, their sizes, and locations. Tumor detection is a crucial initial step in the research, as inaccuracies in the detection phase can lead to unreliable segmentation outcomes. The serial combination of YOLOv3, YOLOv4, and UNet in analyzing the testing data could provide great results in detection and segmentation. For the training section, model 3 is chosen as the best model for YOLOv3-UNet, while model 2 is the best model for YOLOv4-UNet based on the minimum loss and maximum accuracy criteria. These models were then performed to segment the testing dataset.

In the testing section, despite using the original data from Dr. Soetomo, this study adds more challenges by giving data augmentation. The original images added by Gaussian noise and speckle noise are generated as the addition of testing data. The results show that YOLOv3-UNet is the best model in segmenting the original images and Gaussian noisy images with an average CCR of about 97%. Yolov4-UNet, on the other hand, provides results that are not satisfactory. This model's average CCR is 95% for original images and 93% for speckle noisy images. The ROI of the image with Gaussian noise did not recognized by YOLOv4-UNet; therefore, the CCR is missing. Even though YOLOv4 could localize the brain tumor area, the bounding box does not provide an opportunity for the UNet to see a contrasting color other than the tumor area. This is possible since, in the bounding box, the UNet only sees almost the same color due to the addition of the noise.

A comparison with related research, Pravitasari [10], which uses the same source testing data, is carried out to provide a better comparison. With VGG16-UNet segmentation, the average CCR value is 95.69%. The proposed model provides a higher value of CCR, about 97%.

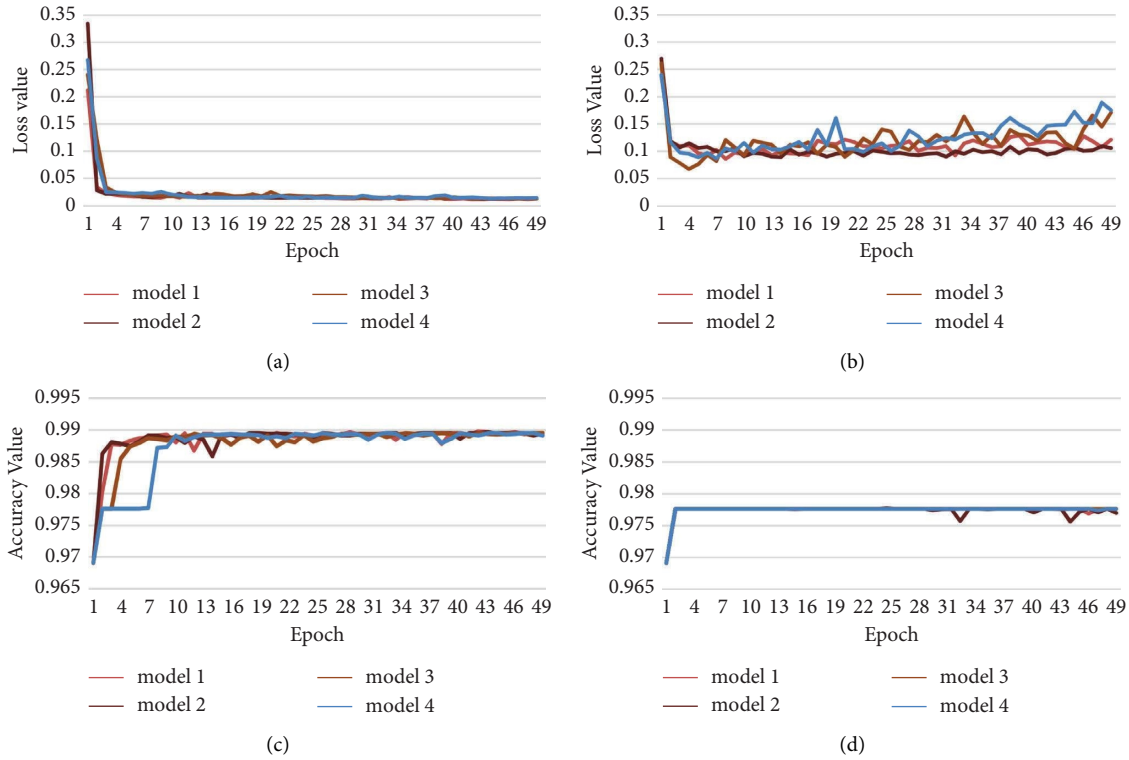


Figure 6: Loss with (a) YOLOv3 and (b) YOLOv4 and accuracy with (c) YOLOv3 and (d) YOLOv4 for each epoch.

Table 5: The comparison of loss and accuracy for each scenario.

	YOLOv3-UNet				YOLOv4-UNet			
	Training		Validation		Training		Validation	
	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy
Model 1	0.0103 (2)	0.9885 (4)	0.0126 (2)	0.9894 (3)	0.0553 (2)	0.9743 (3)	0.1211 (2)	**0.9776 (1)**
Model 2	**0.0096 (1)**	**0.9899 (1)**	0.0135 (3)	0.9895 (2)	0.0587 (4)	0.9740 (4)	**0.1059 (1)**	0.9770 (2)
Model 3	0.0111 (3)	0.9893 (2)	**0.0124 (1)**	**0.9896 (1)**	0.0576 (3)	0.9762 (2)	0.1706 (3)	**0.9776 (1)**
Model 4	0.0120 (4)	0.9887 (3)	0.0140 (4)	0.9891 (4)	**0.0541 (1)**	**0.9777 (1)**	0.1754 (4)	**0.9776 (1)**

(1/2/3/4) indicates the rank of lost and accuracy; (1) is the first rank and so on. Bold value indicates the optimum number of loss and accuracy.

Moreover, as in Table 1, the UNet-VGG16 did not include automatic detection and only provided semantic segmentation. Our proposed model has more advantages in automatic detection and segmentation, which eliminate the manual time-consuming process. Another comparison was conducted between our proposed model and mask R-CNN [15, 16], which have the same implementation to classify individual objects and localize them using bounding boxes and semantic segmentation. The results of comparisons with mask R-CNN are also examined in this study in order to determine if the proposed model offers a higher metric of evaluation for both noisy images and original images. Table 6 shows that mask R-CNN produces a CCR value less than YOLOv3-UNet. Therefore, the proposed model is considered better than the mask R-CNN.

A high CCR value holds positive implications for the detection and segmentation processes. While CCR is employed to measure the accuracy of segmentation outcomes, in this study, segmentation is performed based on the localization results of tumor images, which are cropped from the YOLO bounding box outputs. As a result, the segmentation results are closely tied to the detection outcomes from the preceding phase, owing to the fact that the proposed model follows a serial configuration of detection and

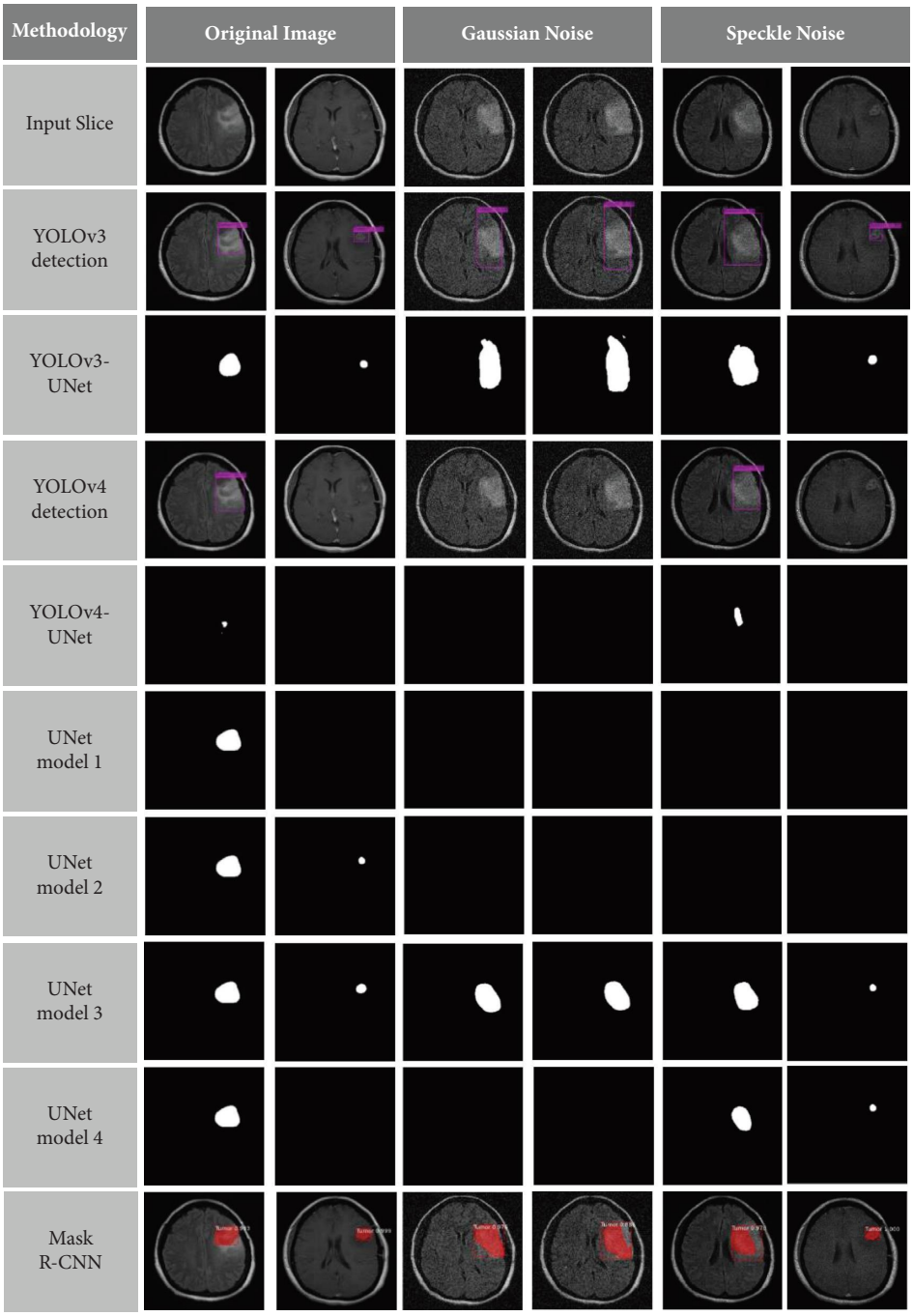


Figure 7: Some of segmentation results for YOLOv3-UNet, YOLOv4-UNet, UNet in four scenarios, and mask R-CNN.

Table 6: The comparison of average CCR for each methodology.

Methodology	Original image	Gaussian noise	Speckle noise
YOLOv3-UNet	*0.978519	*0.973193	0.973546
YOLOv4-UNet	0.958206	—	0.936050
UNet model 1	0.970946	0.959026	0.969738
UNet model 2	0.969377	0.959026	0.969678
UNet model 3	0.975070	0.965331	*0.976370
UNet model 4	0.973683	0.959354	0.974498
Mask R-CNN	0.964181	0.964679	0.964409

*the maximum value of CCR between all methods.

segmentation. A high CCR value indicates that the proposed model excels in accurately recognizing tumors, a crucial aspect in the medical field for effectively distinguishing tumor regions from healthy brain tissue.

6. Conclusions

The study introduces an approach by combining the CNN-FCN methodology, specifically the YOLO-UNet architecture. This combination enables automatic detection and segmentation, which is distinct from previous methods that treated detection and segmentation separately. YOLOv3 and YOLOv4 are chosen for their strong performance and speed in tumor detection. To overcome UNet architecture complexity, the study also explores four scenarios and selects the best model based on high accuracy and low loss values. Tumor detection is challenging due to varying tumor sizes and locations. The YOLOv3 and YOLOv4 are capable of addressing different tumor types, sizes, and locations, crucial for accurate detection.

YOLOv3-UNet is identified as the superior model for segmenting original and Gaussian noisy images, achieving an average correct classification rate (CCR) of about 97%. YOLOv4-UNet performs less satisfactorily, with an average CCR of 95% for original images and 93% for speckle noisy images. The limitation of YOLOv4-UNet in recognizing regions with contrasting colors is noted. A comparison is made with VGG16-UNet and mask R-CNN, which share similar implementation goals. The proposed model's advantages lie in automatic detection and segmentation and eliminating manual efforts of preprocessing. Moreover, the proposed model also shows better results in terms of CCR, further establishing its efficacy.

The study's limitation lies in the localization during the tumor detection process. The bounding box image may not always be the best-suited material for segmentation, especially when the data contain significant noise. Additionally, the research relies on a limited YOLO version, primarily due to the constrained computational specifications of the local hardware. There is potential for improvement and further research in automatic classification and segmentation, possibly through the adoption of updated methods that yield better performance.

Data Availability

The brain tumor data used to support the findings of this study have not been made available because it is ethical to protect patients at Dr. Soetomo Hospital.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are grateful to the Directorate for Research and Community Service (DRPM) Ministry of Research, Technology, and Higher Education Indonesia which supports this research under PT research grant no. 1311/PKS/ITS/2020. The authors also thankful to the Research Center for Artificial Intelligence and Big Data UNPAD for the support.

References

[1] Global Cancer Observatory, "360-indonesia-fact-sheets," 2019, http://gco.iarc.fr/today/data/factsheets/populations/360-indonesia-fact-sheets.pdf.

[2] S. Deepak and P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," *Computers in Biology and Medicine*, vol. 111, Article ID 103345, 2019.

[3] M. Hu, H. Lin, Z. Fan et al., "Learning to recognize chest-xray images faster and more efficiently based on multi-kernel depthwise convolution," *IEEE Access*, vol. 8, pp. 37265–37274, 2020.

[4] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Landmark-based deep multi-instance learning for brain disease diagnosis," *Medical Image Analysis*, vol. 43, pp. 157–168, 2018.

[5] H. I. Suk, S. W. Lee, D. Shen, and Alzheimer's Disease Neuroimaging Initiative, "Deep ensemble learning of sparse regression models for brain disease diagnosis," *Medical Image Analysis*, vol. 37, pp. 101–113, 2017.

[6] Y. Xue, R. Zhang, Y. Deng, K. Chen, and T. Jiang, "A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis," *PLoS One*, vol. 12, no. 6, Article ID e0178992, 2017.

[7] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[8] M. F. Safdar, S. Kobaisi, and F. T. Zahra, "A comparative analysis of data augmentation approaches for magnetic resonance imaging (MRI) scan images of brain tumor," *Acta Informatica Medica*, vol. 28, no. 1, p. 29, 2020.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Munich, Germany, October 2015.

[10] A. A. Pravitasari, N. Iriawan, M. Almuhayar et al., "UNet-VGG16 with transfer learning for MRI-based brain tumor segmentation," *Telkomnika*, vol. 18, no. 3, pp. 1310–1318, 2020.

[11] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "Nas-unet: Neural architecture search for medical image segmentation," *IEEE Access*, vol. 7, pp. 44247–44257, 2019.

[12] J. Li, X. Lin, H. Che, H. Li, and X. Qian, "Probability map guided bi-directional recurrent UNet for pancreas segmentation," 2019, https://arxiv.org/abs/1903.00923.

[13] M. Islam, V. S. Vibashan, V. J. M. Jose, N. Wijethilake, U. Utkarsh, and H. Ren, "Brain tumor segmentation and survival prediction using 3D attention UNet," in *Proceedings of the International MICCAI Brainlesion Workshop*, pp. 262–272, Shenzhen, China, October 2019.

[14] S. Bisa, *Brain tumor detection and segmentation using R-CNN*, Ph.D. thesis, Lamar University, Beaumont, 2020.

[15] M. Süleyman Yıldırım and E. Dandıl, "Automatic detection of multiple sclerosis lesions using Mask R-CNN on magnetic resonance scans," *IET Image Processing*, vol. 14, no. 16, pp. 4277–4290, 2020.

[16] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, Venice, Italy, October 2017.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings*

of the IEEE conference on computer vision and pattern recognition, pp. 779–788, Las Vegas, NV, USA, June 2016.

[18] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525, Honolulu, HI, USA, July 2017.

[19] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.

[20] S. Bhuvaji, A. Kadam, P. Bhumkar, S. Dedge, and S. Kanchan, *Brain tumor classification (MRI) [data set]*, Kaggle, San Francisco, CA, USA, 2020.

[21] B. V. Somasundaran, R. Soundararajan, and S. Biswas, "Robust image retrieval by cascading a deep quality assessment net-work," *Signal Processing: Image Communication*, vol. 80, Article ID 115652, 2020.

[22] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning," https://d2l.ai.2019.

[23] A. A. Pravitasari, N. I. Nirmalasari, N. Iriawan, K. Fithriasari, S. W. Purnami, and W. Ferriastuti, "Bayesian spatially constrained fernandez-steel skew normal mixture model for MRI-based brain tumor segmentation," in *Proceedings of the 2nd In-ternational Conference on Science, Mathematics, Environment, and Education, Solo*, Surakarta, Indonesia, July 2019.

[24] F. Ren, C. Yang, Q. Qiu et al., "Exploiting discriminative regions of brain slices based on 2D CNNs for Alzheimer's disease classification," *IEEE Access*, vol. 7, pp. 181423–181433, 2019.