Journal of Naval Sciences and Engineering
2024, Vol. 20, No. 2, pp. 191-211
Computer Engineering/Bilgisayar Mühendisliği

DOI: 10.56850/jnse.1507140

RESEARCH ARTICLE

**An ethical committee approval and/or legal/special permission has been required within the scope of this study.*

SEGMENTATION OF THYROID NODULES ON ULTRASOUND IMAGES

Burcu BEKTAS GUNES[1*]
Ruya SAMLI[2]
Mahmut Bilal DOGAN[3]
Duzgun YILDIRIM[4]

[1]*National Defence University, Turkish Naval Academy, Department of Computer Engineering, Istanbul, Turkey,*
burcubektasgunes@gmail.com

[2]*Istanbul University-Cerrahpasa, Department of Computer Engineering, Istanbul, Turkey,*
ruyasamli@iuc.edu.tr

[3]*Istanbul Medeniyet University, Department of Radiology, Istanbul, Turkey,*
mahmutbilaldogan@gmail.com

[4]*Acidabem Kozyatagi and Taksim Hospitals, Departments of Radiology, Istanbul, Turkey,*
drduzgunyildirim@gmail.com

Received: 29.06.2024 **Accepted: 23.10.2024**

ABSTRACT

The increasing prevalence of thyroid cancer in our country and globally has led to the development of various computer-aided studies for its detection, contributing significantly to the literature. Artificial intelligence and image processing are particularly prominent methods in this field due to their non-invasive nature, accessibility, and ability to provide valuable information about the morphological characteristics of nodules. In recent years, segmentation algorithms in medical imaging have garnered substantial interest for their potential to enhance diagnostic accuracy. Accurate segmentation of thyroid nodules is a critical first step in the development of AI-assisted clinical decision support systems for the detection and diagnosis of thyroid cancer.

In this study, innovative methods were employed to detect thyroid nodules. A dice score of 79% was achieved in instance segmentation using the YOLOv5-Small algorithm when doppler images were excluded, while a dice score of 91% was obtained using the YOLOv5-Large algorithm on a dataset that included doppler images. In semantic segmentation, the Attention Unet++ and Manet algorithms achieved a dice score of 89% when doppler images were excluded, and 91% when they were included. These results demonstrate that images typically excluded by physicians could potentially offer better outcomes in computerized image processing.

Keywords: *Thyroid Nodule Segmentation, Artificial Intelligence, Instance Segmentation, Semantic Segmentation.*

ULTRASON GÖRÜNTÜLERİNDE TİROİD NODÜLLERİNİN SEGMENTASYONU

ÖZ

Ülkemizde ve dünyada tiroid kanser miktarının yaygınlaşması neticesinde tiroid kanserlerinin tespit edilebilmesi için bilgisayar destekli farklı çalışmaların yapılması literatüre önemli bir katkı sağlamaktadır. Özellikle yapay zeka ve görüntü işleme konuları bu alanda sıklıkla kullanılan bir yöntemdir. Bunun nedeni, girişimsel olmayan yapısı, erişilebilirliği ve nodüllerin morfolojik özellikleri hakkında değerli bilgiler sağlayabilmesidir. Son yıllarda, tıbbi görüntülemede segmentasyon algoritmaları, tanısal doğruluğu artırma potansiyelleri nedeniyle büyük ilgi görmüştür. Tiroid nodüllerinin doğru segmentasyonu, yapay zeka destekli klinik karar destek sistemlerinin tiroid kanserinin tespiti ve teşhisi için geliştirilmesinde kritik bir ilk adımdır.

Bu çalışmada tiroid nodüllerinin tespit edilebilmesi için yenilikçi bazı yöntemler kullanılmıştır. Örnek segmentasyonunda YOLOv5-Small algoritması ile doppler görüntüleri hariç tutulduğunda %79 dice skoru sağlanmıştır, sonrasında doppler görüntülerini içeren veri setinde YOLOv5-Large algoritması ile %91 test dice skoru elde edilmiştir. Semantik segmentasyonda Attention Unet++ ve Manet algoritması kullanılarak, doppler görüntüleri hariç tutulduğunda %89 test dice skoru elde edilirken, doppler görüntülerini içeren veri setinde %91 test dice skoruna ulaşılmıştır. Böylece normal şartlarda hekimler tarafından hariç tutulan görüntülerin de bilgisayarlı görüntü işleme sürecinde daha yüksek sonuçlar sunabileceği gösterilmiştir.

Anahtar Kelimeler: *Tiroid Nodül Segmentasyonu, Yapay Zeka, Örnek Segmentasyonu, Semantik Segmentasyon.*

1. INTRODUCTION

The thyroid is the part of the endocrine system that is located in the front of the throat region and produces, stores, and secretes hormones. During routine checks or based on patient complaints, abnormal nodules of different structures and sizes can sometimes form within the thyroid tissue, and these nodules can lead to cancer in advanced stages. Thyroid nodules are a common medical condition affecting a significant portion of the population, with prevalence rates ranging from 19% to 68% (Demetriou et al., 2023).

Clinical evaluations consider other risk factors such as family history of thyroid disease, the patient's age, gender, radiation exposure, and the presence of any endocrine disorders. Palpation is the first method employed in assessing the disease at this stage, where physical findings such as the presence, size, and hardness of the nodule in the neck area are checked. Subsequently, a blood test is conducted to examine certain hormones (TSH, freeT3, freeT4 and some relevant autoantibodies), and if deemed necessary, the patient is referred to the radiology department. The most useful imaging technique is ultrasonography. According to an epidemiological study, thyroid nodules are frequently detected in up to 70% of ultrasound examinations; however, only 5-15% of sonographically detected nodules are malignant (Eloy et al., 2022). Therefore, the primary clinical challenge is to reliably distinguish malignant nodules that require surgical treatment from the majority of benign nodules that do not necessitate surgery. To reduce unnecessary biopsies while detecting clinically significant cancers, various risk classification systems, such as the American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TIRADS), have been proposed (Tessler et al., 2017). The nodule detected on ultrasonography is evaluated according to TIRADS criteria and a risk stratification is made. And a biopsy sample is taken from nodules identified as high risk under ultrasound guidance and examined microscopically. According to cytopathological examination findings, it is determined whether the nodule is benign or malignant.

While TIRADS systems have reduced biopsy rates and increased specificity, a significant number of false-positive biopsies, reported at 49-56%, persist (Hoang, Middleton, Farjat, Langer, et al., 2018; Hoang, Middleton, Farjat, Teefey, et al., 2018; Yamashita et al., 2022). Additionally, these systems rely on the interpretation of qualitative imaging features and are subject to intra- and inter-observer variability even among expert radiologists (Kappa=0.519) (Hoang, Middleton, Farjat, Teefey, et al., 2018).

As a result, the subjectivity and variability inherent in human interpretation of the diagnostic tools have prompted the exploration of AI as a potential solution to enhance the accuracy and consistency of thyroid nodule risk. The accurate segmentation of thyroid nodules is a crucial and first step in the development of AI-powered clinical decision support systems for the detection and diagnosis of thyroid cancer.

This study, analyzes and compared the performance of semantic and instance segmentation algorithms for the automatic segmentation of thyroid nodules, using a unique thyroid ultrasound dataset. During the implementation of the study, multiple models were tested, and in this manuscript, we present the results of the three algorithms that provided the best performance on our dataset. The Attention Unet++ and MAnet models incorporate attention mechanisms that allow for enhanced feature extraction and improved localization of nodules in complex ultrasound images. The YOLOv5 model, real-time instance segmentation algorithm, is highly effective in detecting multiple objects within an image, which aligns with the need for accurate nodule detection in multiple-frame ultrasound data. Segmentation experiments were conducted across various dataset versions, and the models performed better when doppler images were included. Even though physicians typically exclude doppler images from consideration, it was shown that these images improved segmentation performance, suggesting that doppler images can be beneficial in this context.

2. RELATED WORK

The segmentation methods most commonly employed for thyroid nodules can be grouped into three main categories: active contour, feature extraction and classification by traditional machine learning algorithms, and deep learning (Chen et al., 2020). While traditional segmentation methods have made some progress in computer-aided diagnosis, they are influenced by subjective factors in feature selection and parameter setting. Therefore, in order to meet the clinical needs of the field, it is crucial to achieve automatic segmentation of thyroid nodules with higher accuracy. Recently, state-of-the-art algorithms such as Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) have demonstrated considerable potential in the segmentation of medical ultrasound images (Chen et al., 2020). For instance, Zhou et al. proposed a hierarchical classification system including ResNet and Conditional Variable Autoencoder (CVAE). The proposed system can identify benign and malignant thyroid nodules with an accuracy of 0.892 and an AUC of 0.887, according to testing on real patient (Zhou et al., 2022). Aytac et al. used CNN techniques to segment thyroid nodules based on TIRADS categories, achieving an average accuracy of 91.38% (Aytaç et al., 2021). Buda et al. proposed and assessed two methods for deep learning-based thyroid nodule segmentation that incorporate the use of calipers present in the images. The first method utilized approximate nodule masks generated from the calipers, while the second method integrated manual annotations with automatic guidance provided by the calipers. The method leveraging caliper guidance improved the Dice Similarity Coefficient (DSC) of nodule segmentation from 85.1% to 93.1% (Buda et al., 2020). Abdolali et al. the core of the proposed method is a deep learning framework based on the multi-task mask R-CNN model. They developed a regularization loss function that prioritizes detection over segmentation. The experimental results indicate that their proposed method outperforms Faster R-CNN and Mask R-CNN in thyroid nodule detection (Abdolali et al., 2020). Gong et al. developed an innovative multi-task learning framework that concurrently learns the nodule size, gland position, and nodule position. Additionally, they proposed an adaptive gland region feature enhancement module to fully utilize prior knowledge of the thyroid gland (Gong et al., 2023). Kunapinun et al. demonstrated a new algorithm for segmenting thyroid nodules in ultrasound images. The algorithm combined traditional supervised semantic segmentation with unsupervised learning using GANs (Kunapinun et al., 2023). In a study conducted by Gokmen Inan et al., the thyroid nodule was segmented using U-Net architectures, namely ResUNet and ResUNet++. These were integrated with feature extraction and upsampling with dropout operations, which were employed to prevent overfitting (Gökmen Inan et al., 2024).

3. MATERIALS AND METHOD

3.1. Dataset

With the 2022/0503 approval number of the Ethics Committee Ethics Committee of Istanbul Medeniyet University on 07.09.2022, images of registered patients who underwent thyroid biopsy between 2018-2020 were retrieved from the picture archiving and communication systems (PACS) system of Istanbul Medeniyet University Göztepe Prof. Dr. Süleyman Yalçın City Hospital.

The images were labeled by a radiology specialist with 10 years of head and neck radiology experience on the ango.ai platform. During this labeling process, as shown in Figure 1, the TIRADS classification and the nodule shape characteristics were labeled.

The ango.ai platform provides a JSON output format for the labeled images. This JSON file was executed with the help of a Python script to create images and masks for segmentation. The You Only Look Once (YOLO) and Common Objects in Context (COCO) formats were obtained. The primary reason for generating both formats was to store the masks of the labeled images in a standard format. This standardization ensured consistency across the dataset and facilitated compatibility with different segmentation models. As the images were labeled, they were versioned, and training trials were conducted.

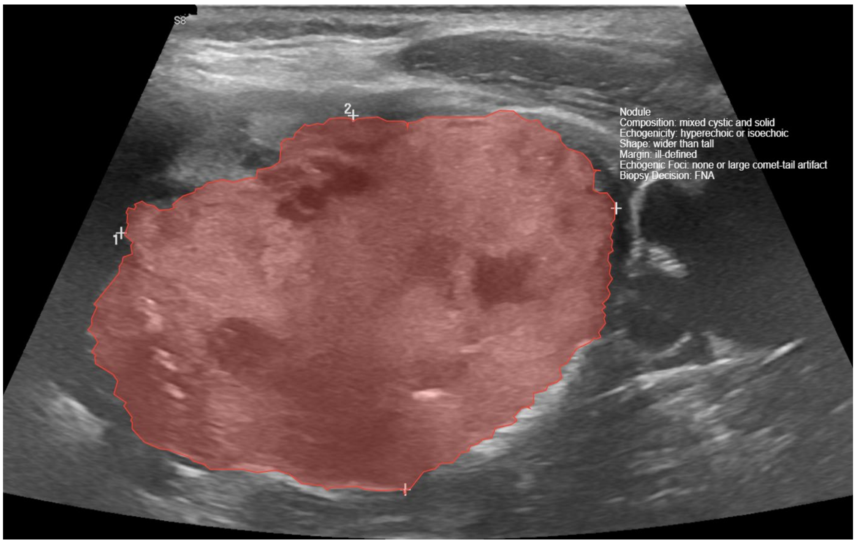


Figure 1. *Standard for image labeling*

Some images were labeled as invalid or containing artifacts by the physicians. These images (n=829) were identified and excluded from the study, creating a dataset labeled as Version-1 (V1). The suitability of the V1 images for artificial intelligence was analyzed by developers. In 197 images, it was detected that the doppler-elastography feature was enabled. These images (n=197) were identified and excluded from the study, and a new dataset labeled as Version-2 (V2) was created. The exclusion of the 197 images when creating V2 was done intentionally to compare the model's performance with and without doppler images. This comparison aimed to evaluate whether doppler images offered significant improvements in segmentation, which they did. Figure 2 provides the characteristics of the datasets obtained for the segmentation study.

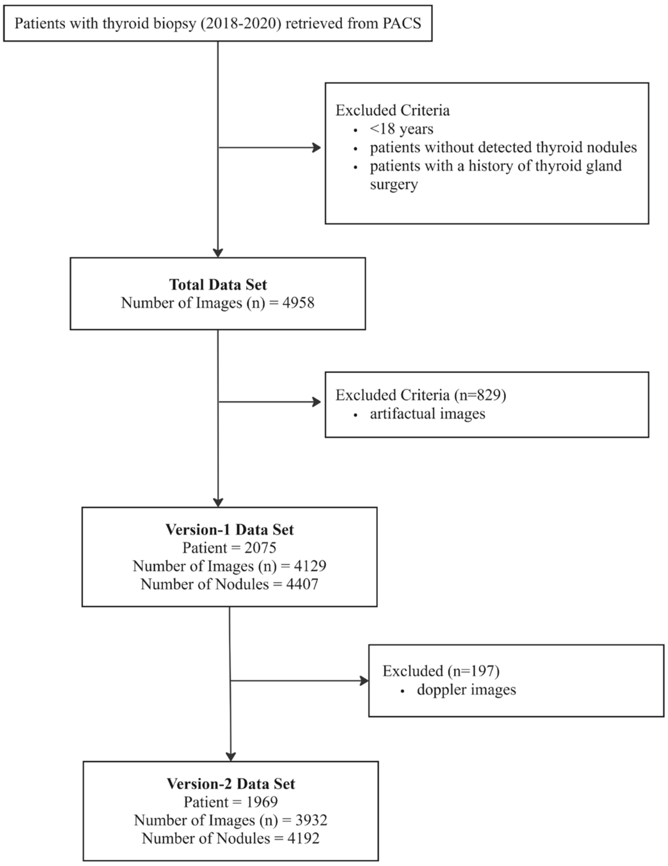
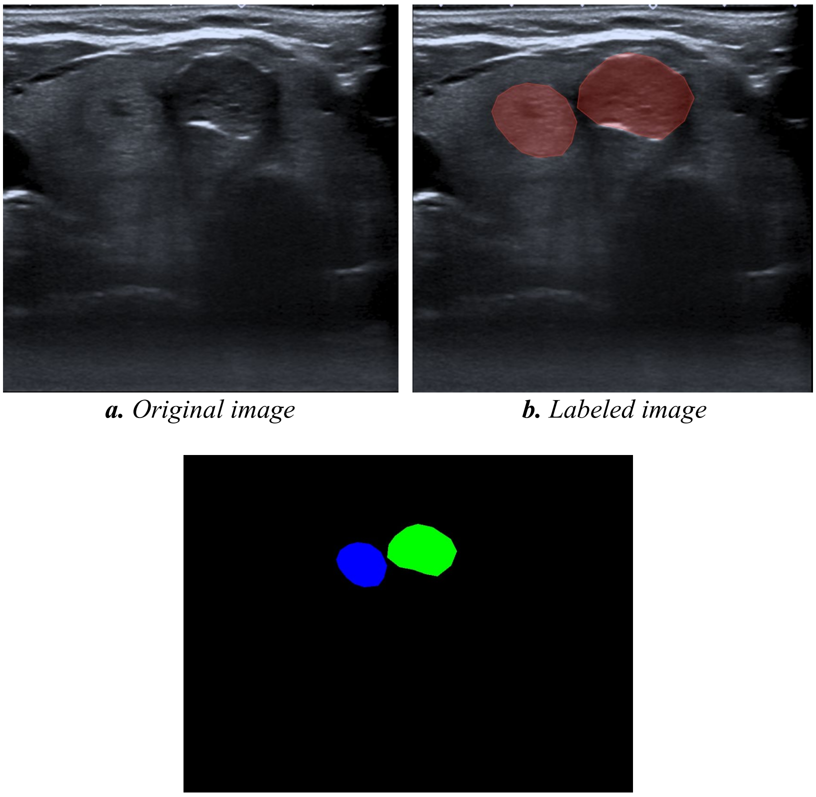


Figure 2. *Dataset versions created for segmentation*

During the ultrasound imaging, a patient may have multiple images. It is also possible for each image to contain more than one nodule. In this study, since the segmentation process will be performed for each nodule, each nodule has been labeled and image masks have been saved.

For example, the patient indicated in Figure 3.a has 2 nodules, and these nodules are labeled as shown in Figure 3.b. The masks for these 2 nodules are saved as shown in Figures 3.c.



a. *Original image* ***b.*** *Labeled image*

c. *Mask of the nodule*

Figure 3. *Nodule masks obtained from labeled images*

3.2. Data Pre-processing

The steps performed during the data preprocessing phase can be summarized as follows:

- Conversion of Raw DICOM Images: Using the PyDicom library, raw DICOM images were converted to PNG formats to facilitate further image processing and analysis. This conversion allows for better compatibility with various image processing tools. Given that the DICOM images are 8-bit, no compression was applied, ensuring that no information was lost during this stage.

- Normalization of Images: The images are 8-bit and grayscale. To normalize the images, each pixel value was scaled to the range [0, 1], ensuring uniformity and enhancing the quality of subsequent analyses.
- Extraction of Nodule Regions: Using a customized JSON script, the regions of the nodules drawn as polygons, also known as mask images, were extracted from the JSON data. These regions, were essential for accurately identifying and analyzing nodule areas.
- Conversion of Masks: The extracted masks were converted into YOLO and COCO formats to facilitate compatibility with various object detection frameworks and enhance the efficiency of model training.

3.3. Semantic Segmentation

Image segmentation is essential in medical imaging, as it plays a vital role in disease analysis and diagnosis. Accurate identification of the region of interest (ROI) within a sample is a key step in any feature segmentation process. Semantic-based image segmentation involves pixel-level classification, enabling the detection of biological structures and the quantification of their morphology. Moreover, it facilitates the quantitative capture of object shapes and supports high-resolution spatial statistical analysis (Qureshi et al., 2023).

In the literature, the most widely used U-Net-based models, incorporating the latest advancements in medical image segmentation, such as Attention UNet++ and Multi-scale Attention Net (MAnet), were tested. UNET, a CNN framework, features a simple U-shaped encoder-decoder architecture. Despite being trained on a small dataset, it delivers precise segmentation results (Hettihewa et al., 2023).

3.4. Instance Segmentation with YOLO

The goal of instance segmentation is to accurately identify and outline each instance of a class within an image (Sharma et al., 2022). YOLOv5 is a state-of-the-art single-stage object detection network, known for being lightweight and efficient. Its compact size makes it easier to train and quicker to generate predictions, making it particularly suitable for real-time applications (Jocher, 2020; Yang et al., 2024). In this study, different versions of YOLOv5 were tested, with the YOLOv5-Large model showing the best performance in terms of speed and accuracy. As a result, the comparison was focused on the performance variations between these YOLOv5 models.

Before starting the segmentation process, data pre-processing steps were performed using the JSON format containing image labels. Outputs were generated from the JSON file, which included nodule image and mask images, according to YOLO and COCO formats. After this stage, the images and masks were included in the training phase based on instance segmentation.

3.5. Evaluation Metrics

To evaluate the performance of the proposed method, several standard metrics were used, including Dice score, Precision, Recall. These metrics define the following parameters: False Negative (FN) represents a positive sample misclassified as negative, while True Negative (TN) is correctly classified negative. False Positive (FP) is a negative sample misclassified as positive and True Positive (TP) indicates correctly classified as positive.

3.5.1. Dice Score

As stated Formula 1 (Zou et al., 2004), the Dice score, which measures the overlap between the predicted segmentation and the ground truth, is a commonly used metric for evaluating segmentation performance.

$$\frac{2TP}{2TP + FP + FN} \tag{1}$$

3.5.2. Precision

As stated Formula 2 (Hicks et al., 2022), the accuracy of the detected objects, indicating how many detections were correct (Ultralytics, 2024).

$$\frac{TP}{TP + FP} \tag{2}$$

3.5.3. Recall

As stated Formula 3 (Hicks et al., 2022), the ability of the model to identify all instances of objects in the images (Ultralytics, 2024).

$$\frac{TP}{TP + FN} \tag{3}$$

3.5.4. Mean Average Precision (mAP_0.5)

The mean Average Precision (mAP) is the most commonly used evaluation metric. Precision is calculated based on the Intersection over Union (IoU), which measures the ratio between the area of overlap and the area of union between the predicted bounding box and the ground truth. A threshold is applied to determine whether a detection is correct. When the IoU exceeds the threshold, it is classified as a True Positive, while an IoU below the threshold is classified as a False Positive. If the model fails to detect an object present in the ground truth, it is considered a False Negative (Zaidi et al., 2022).

mAP at IoU threshold of 0.5 was used to evaluate the performance of the nodule detection task.

4. RESULT

The datasets used for the segmentation section were divided into 80% for training, 15% for validation, and 5% for testing. Details about the top four algorithms that produced the best results are provided.

In this study, data augmentation techniques were not applied. The primary reason is that the thyroid nodule images used in our research are ultrasound images, and these images can be sensitive to changes in orientation, scale, or position due to their inherent nature. Applying transformations such as rotation or scaling may distort the anatomical structures of the nodules, potentially impacting the diagnostic accuracy. Therefore, the original format of the data was maintained to ensure the preservation of clinically relevant information. Despite the limited size of the dataset, it was determined that the existing data provided sufficient diversity to optimize model performance.

Additionally, dropout and weight decay regularization techniques were not employed during training. This decision was based on the satisfactory accuracy levels observed throughout the training process, which did not indicate a need for further regularization. The absence of overfitting in our models, as evidenced by consistent performance across training and validation datasets, further justified the exclusion of these regularization methods.

The semantic segmentation algorithms were trained using different hyperparameters. Table 1 presents the hyperparameters of the Attention Unet++ and MAnet semantic segmentation algorithms with the best results, while Table 2 shows their performance details. For the algorithms that produced the best results, focal loss (FL) was chosen as the loss function, and stochastic gradient descent (SGD) was selected as the optimizer. The backbone can be of different types; in this work, ResNext50 were considered. While some hyperparameters were taken from previous study, we also performed optimization experiments to fine-tune the settings for our specific dataset. For instance, the learning rate and batch size were adjusted iteratively to find a balance between training speed and model performance. Additionally, the number of epochs was selected to prevent overfitting while ensuring the model reached optimal accuracy.

Table 1. *Details of semantic segmentation hyperparameters*

Model Version	**Model**	**Schedula r**	**Initial LR**	**Epochs**	**Batch Size**	**Parameter (M)**	**Image Size**	**GPU**
1	Attention Unet++	ReduceLR OnPlateau	0.0003	50	16	2.0	720	Tesla K80
2	Attention Unet++	ReduceLR OnPlateau	0.0003	30	16	47.9	720	Tesla K80
3	MAnet	ReduceLR OnPlateau	0.0003	50	16	7.6	720	Tesla K80
4	MAnet	ReduceLR OnPlateau	0.0003	50	16	22.0	720	Tesla K80

LR: Learning Rate

Table 2. *Dice and AUROC metrics for the Semantic segmentation models*

Model	Data Set Version	Model Version	Test Dice Score	Test AUROC
Attention Unet++	**1**	**1**	**91%**	**99%**
MAnet	1	3	91%	99%
MAnet	1	4	89%	99%
Attention Unet++	**2**	**1**	**89%**	**99%**

UNet-based models, including Attention Unet++ and MAnet, which represent the most widely utilized and advanced approaches in medical image segmentation, were evaluated. These two models, incorporating the ResNext50 32x4D Encoder network, achieved the highest test dice score of 91% for the V1 data set. Figure 4 illustrates the test output of Attention Unet++ model on a sample image. In the figure, the color coding is as follows: Green denotes the ground truth, Red indicates the prediction, and Yellow represents the overlap.

The instance segmentation algorithms were trained using different hyperparameters. Table 3 shows the hyperparameter configurations that yielded the best results, while Table 4 presents the performance details. In the algorithms that give the best results, the loss function is determined as focal loss and image size as 720.

Figure 4. *Attention Unet++ model inference*
Green: Ground truth, Red: predicted segmentation, yellow: overlap between the two

Table 3. *Details of instance segmentation hyperparameters*

Model	**Schedular**	**Initial LR**	**Epochs**	**Batch Size**	**Parameters (M)**	**Optimizer**	**GPU**
YOLOv5 - Nano	Cosine LR Scheduler	0.00001	50	16	2.0	SGD	Tesla K80
YOLOv5 - Small	Cosine LR Scheduler	0.00001	50	16 / 32	7.6	SGD	Tesla K80 / Nvidia P100
YOLOv5 - Medium	Cosine LR Scheduler	0.00001	50	16	22.0	SGD	Tesla K80 / Nvidia P100
YOLOv5 - Large	Cosine LR Scheduler	0.00001	30	16	47.9	SGD	Tesla K80
YOLOv5 - XLarge	Cosine LR Scheduler	0.00001	25	8	88.8	Adam	Tesla K80

Table 4. *Various performance metrics for the Yolov5 models*

Model	DV	Test DS	Valid FL	mAP_ 0.5* Masks	mAP_ 0.5 Boxes	P Masks	P Boxes	Recall Masks	Recall Boxes
YOLOv5 - Nano	1	0.84	0.027	0.87	0.87	0.88	0.87	0.795	0.79
YOLOv5 - Nano	2	0.75	0.027	0.79	0.79	0.75	0.79	0.78	0.78
YOLOv5 - Small	1	0.87	0.028	0.87	0.87	0.89	0.88	0.80	0.80
YOLOv5 - Small	**2**	**0.799**	**0.293**	**0.85**	**0.849**	**0.853**	**0.85**	**0.76**	**0.75**
YOLOv5 - Medium	1	0.90	0.03	0.88	0.88	0.90	0.89	0.80	0.82
YOLOv5 - Medium	2	0.793	0.288	0.84	0.84	0.831	0.825	0.78	0.79
YOLOv5 - Large	**1**	**0.91**	**0.028**	**0.87**	**0.87**	**0.88**	**0.88**	**0.81**	**0.81**
YOLOv5 - Large	2	0.76	0.028	0.75	0.76	0.75	0.78	0.71	0.72

DV: Dataset Version
DS: Dice Score
P: Precision

In addition to semantic segmentation, a real-time image segmentation model based on YOLOv5 was presented. The YOLOv5-Large model provided the best result for the test dice score, achieving 91%. Figure 5 shows the prediction results of the YOLOv5-Large model on the test image.

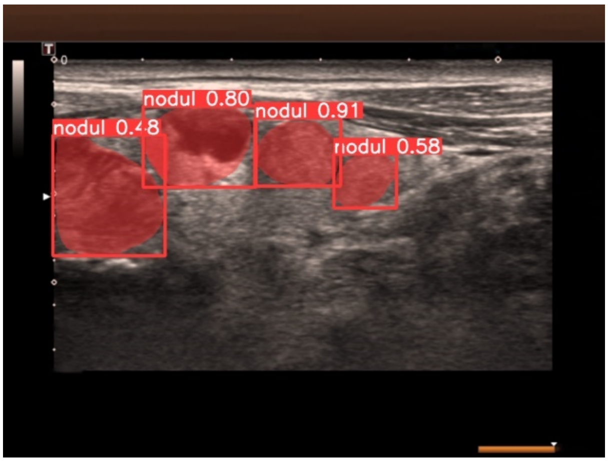


Figure 5. *YOLOv5-Large inference of multi-nodule segmentation*

This study highlights the growing importance of computer-aided methods in the detection of thyroid nodules, with a particular focus on artificial intelligence and image processing techniques. Our findings emphasize the critical role of accurate segmentation in developing AI-assisted clinical decision support systems for thyroid cancer detection. By comparing various segmentation algorithms, including YOLOv5 and Attention Unet++, we demonstrated that incorporating doppler images—often excluded by physicians—can significantly improve segmentation performance. Specifically, the YOLOv5-Large and Attention Unet++ algorithms achieved higher dice scores when doppler images were included, underscoring the potential benefits of utilizing these images in computerized image processing. These results contribute to the ongoing efforts to enhance diagnostic accuracy in thyroid cancer detection, paving the way for more effective and reliable AI-driven healthcare solutions.

One limitation of this study is the use of a single institution's dataset, which may introduce a bias and limit the generalizability of the findings to other clinical settings. Additionally, the algorithms were evaluated only on static images, without considering temporal variations that might be present in sequential imaging. Moreover, the study did not account for inter-observer variability in the manual labeling of nodules, which could influence the reported segmentation accuracy. Future studies should incorporate more diverse datasets and consider dynamic imaging and inter-observer consistency to strengthen the robustness of the results.

ACKNOWLEDGEMENT

This study was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under the 2214-A International PhD Research Fellowship Program with the number 1059B142100676 and the 2211-C Domestic PhD Fellowship Program for Priority Areas with the number 1649B032001140.

REFERENCES

Abdolali, F., Kapur, J., Jaremko, J. L., Noga, M., Hareendranathan, A. R., & Punithakumar, K. (2020). Automated thyroid nodule detection from ultrasound imaging using deep convolutional neural networks. *Computers in Biology and Medicine, 122,* 103871. https://doi.org/10.1016/j.compbiomed.2020.103871

Aytaç, Z., Iseri, İ., & Dandıl, B. (2021). Derin Öğrenme Kullanarak Tiroid Kanseri Teşhisi. *Avrupa Bilim ve Teknoloji Dergisi*, (29), 292-298.

Buda, M., Wildman-Tobriner, B., Castor, K., Hoang, J. K., & Mazurowski, M. A. (2020). Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images. *Ultrasound in medicine & biology*, *46*(2), 415-421.

Chen, J., You, H., & Li, K. (2020). A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images. *Computer methods and programs in biomedicine*, *185*, 105329.

Demetriou, E., Fokou, M., Frangos, S., Papageorgis, P., Economides, P. A., & Economides, A. (2023). Thyroid nodules and obesity. *Life*, *13*(6), 1292.

Eloy, C., Russ, G., Suciu, V., Johnson, S. J., Rossi, E. D., Pantanowitz, L., & Vielh, P. (2022). Preoperative diagnosis of thyroid nodules: An integrated multidisciplinary approach. *Cancer Cytopathology*, *130*(5), 320-325.

Inan, N. G., Kocadağlı, O., Yıldırım, D., Meşe, İ., & Kovan, Ö. (2024). Multi-class classification of thyroid nodules from automatic segmented ultrasound images: Hybrid ResNet based UNet convolutional neural network approach. *Computer Methods and Programs in Biomedicine*, *243*, 107921.

Gong, H., Chen, J., Chen, G., Li, H., Li, G., & Chen, F. (2023). Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in biology and medicine*, *155*, 106389.

Hettihewa, K., Kobchaisawat, T., Tanpowpong, N., & Chalidabhongse, T. H. (2023). MANet: a multi-attention network for automatic liver tumor segmentation in computed tomography (CT) imaging. *Scientific Reports*, *13*(1), 20098.

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, *12*(1), 5979.

Hoang, J. K., Middleton, W. D., Farjat, A. E., Langer, J. E., Reading, C. C., Teefey, S. A., & Tessler, F. N. (2018). Reduction in thyroid nodule biopsies and improved accuracy with American College of Radiology Thyroid Imaging Reporting and Data System. *Radiology*, *287*(1), 185-193.

Hoang, J. K., Middleton, W. D., Farjat, A. E., Teefey, S. A., Abinanti, N., Boschini, F. J., Bronner, A. J., Dahiya, N., Hertzberg, B. S., Newman, J. R., Scanga, D., Vogler, R. C., & Tessler, F. N. (2018). Interobserver Variability of Sonographic Features Used in the American College of Radiology Thyroid Imaging Reporting and Data System. AJR. *American Journal of Roentgenology, 211*(1), 162–167. https://doi.org/10.2214/AJR.17.19192

Jocher, G. (2020). YOLOv5 by Ultralytics (Version 7.0) [Python]. https://doi.org/10.5281/zenodo.3908559

Kunapinun, A., Dailey, M. N., Songsaeng, D., Parnichkun, M., Keatmanee, C., &Ekpanyapong, M. (2023). Improving GAN Learning Dynamics for Thyroid Nodule Segmentation. *Ultrasound in Medicine & Biology, 49*(2), 416–430. https://doi.org/10.1016/j.ultrasmedbio.2022.09.010

Qureshi, I., Yan, J., Abbas, Q., Shaheed, K., Riaz, A. B., Wahid, A., Khan, M. W. J., & Szczuko, P. (2023). Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion, 90*, 316–352. https://doi.org/10.1016/j.inffus.2022.09.031

Sharma, R., Saqib, M., Lin, C. T., & Blumenstein, M. (2022). A Survey on Object Instance Segmentation. https://opus.lib.uts.edu.au/handle/10453/167620

Tessler, F. N., Middleton, W. D., Grant, E. G., Hoang, J. K., Berland, L. L., Teefey, S. A., Cronan, J. J., Beland, M. D., Desser, T. S., Frates, M. C., Hammers, L. W., Hamper, U. M., Langer, J. E., Reading, C. C., Scoutt, L. M., & Stavros, A. T. (2017). ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. *Journal of the American College of Radiology: JACR, 14*(5), 587–595. https://doi.org/10.1016/j.jacr.2017.01.046

Ultralytics. (2024, June 27). YOLO Performance Metrics. https://docs.ultralytics.com/guides/yolo-performance-metrics

Yamashita, R., Kapoor, T., Alam, M. N., Galimzianova, A., Syed, S. A., Ugur Akdogan, M., Alkim, E., Wentland, A. L., Madhuripan, N., Goff, D., Barbee, V., Sheybani, N. D., Sagreiya, H., Rubin, D. L., & Desser, T. S. (2022). Toward Reduction in False-Positive Thyroid Nodule Biopsies with a Deep Learning-based Risk Stratification System Using US Cine-Clip Images. *Radiology. Artificial Intelligence, 4*(3), e210174. https://doi.org/10.1148/ryai.210174

Yang, D., Xia, J., Li, R., Li, W., Liu, J., Wang, R., Qu, D., & You, J. (2024). Automatic Thyroid Nodule Detection in Ultrasound Imaging With Improved YOLOv5 Neural Network. IEEE Access, 12, 22662–22670. *IEEE Access.* https://doi.org/10.1109/ACCESS.2024.3359367

Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., & Lee, B. (2022). A survey of modern deep learning based object detection models. *Digital Signal Processing, 126*, 103514. https://doi.org/10.1016/j.dsp.2022.103514

Zhou, H., Wang, R., Zhou, M., Fu, P., & Bai, Y. (2022). A Deep Learning-Based Cascade Automatic Classification System for Malignant Thyroid Nodule Recognition in Ultrasound Image. *Journal of Physics: Conference Series*, 2363(1), 012029. https://doi.org/10.1088/1742-6596/2363/1/012029

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M. C., Kaus, M. R., Haker, S. J., Wells, W. M., Jolesz, F. A., & Kikinis, R. (2004). Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. *Academic Radiology, 11*(2), 178–189. https://doi.org/10.1016/S1076-6332(03)00671-8