Development of Hybrid U-Net Models with Pre-trained VGG16 for Thyroid Nodule Segmentation

Elifnur ERDEMİR
Department of Computer Engineering
Başkent University
Ankara, Türkiye
elifnurerdemir@baskent.edu.tr

Çağatay Berke ERDAŞ
Department of Computer Engineering
Başkent University
Ankara, Türkiye
berdas@baskent.edu.tr

***Abstract*—Thyroid nodules are defined as abnormal cell growths that appear as lumps in the thyroid gland. Thyroid nodules are common and detected incidentally due to their asymptomatic nature. Ultrasonography is used to detect thyroid nodules and 7-15% of the nodules detected have the risk of transforming into carcinoma. These carcinomas cause thyroid cancers, but thyroid cancers have a long prognosis and life expectancy after early diagnosis and treatment. Segmentation of thyroid nodules from ultrasound images can improve diagnosis and treatment by accurately identifying regions of interest. This work focuses on developing hybrid U-Net models by incorporating a pre-trained VGG16 model into the encoder of these U-Net architectures, specifically U-Net, U-Net++, and U-Net3+. Experimental results show that these hybrid models outperform their traditional counterparts, achieving higher performance metrics, especially regarding Intersection Over Union and Dice Scores. The findings show that hybrid U-Net models, especially hybrid U-Net3+, offer significant potential to improve thyroid nodule segmentation, with hybrid U-Net3+ achieving the highest Intersection Over Union of 65,404 and a Dice Score of 76,271, which can help in more accurate diagnosis.**

Keywords—Thyroid nodules segmentation; U-Net; UNet++; UNet3+; VGG16; Hybrid U-Net; Hybrid UNet++; Hybrid UNet3+

I. INTRODUCTION

Thyroid nodules are defined as abnormal cell growth that forms a lump within the thyroid gland [1]. Thyroid nodules are clinically difficult to diagnose but are common lesions [2]. Thyroid nodules are mostly asymptomatic and therefore discovered incidentally by palpation [3]. However, examination of the thyroid gland by palpation can be insensitive and inaccurate depending on the skill of the examiner [3]. Therefore, ultrasonography plays an important role in the detection of deeply located and small nodules [2]. The detection rate of thyroid nodules with ultrasound imaging is 20-67%. [1]. Carcinoma develops in 7-15% of these detected nodules and causes thyroid cancers [1, 2]. Thyroid cancers are slow-growing and have a high probability of survival after treatment. Therefore, early diagnosis of nodules is important. [2].

Automatic segmentation of thyroid nodules from ultrasound images can improve and accelerate clinical diagnosis and treatment. Deep learning is a very suitable solution for automatic segmentation. Accurate nodule detection and localization can reduce the workload of doctors and provide rapid diagnosis for patients.

The aim of the study was to segment thyroid nodules using deep learning. First, thyroid nodules were segmented using U-Net, UNet++ and UNet3+, which are models frequently used in medical fields. Then, the VGG16 model trained with thyroid ultrasound images was used in the encoder section of these three U-Net models to create a feature map. Thus, three hybrid U-Net models were obtained. When the results of the experiments were examined, it was revealed that hybrid models gave better results, but the best result was given by hybrid UNet3+.

II. RELATED WORKS

Gong et al. [4] developed a multitask learning framework for thyroid nodule segmentation in their model named TRFE+. This model can learn nodule size estimation, gland location segmentation and nodule location segmentation simultaneously. When they evaluated their model on TN3K dataset, they obtained 92.56 AUC (Accuracy Under Curve), 72.20 F1-Score, 60.47 IoU (Intersection over Union) and 75.37 Dice Score.

Ozcan et al. [5] presented a new segmentation model called Enhanced-TransUNet for the segmentation of thyroid nodules from ultrasound images. They integrated Vision Transformers into the U-Net model. In the TransUNet model, important features were extracted using CNN and then these features were divided into smaller pieces for the Transformer to analyze. After the analysis phase of the Transformer, the model was run with U-Net, using a decoder with upsampling and skip connections to combine the features and produce the final segmented image. As a result, they achieved an IoU score of 70.87 and a Dice Score of 82.92 on the TN3K Dataset with this new model. However, it

979-8-3315-2981-9/24/$31.00 ©2024 IEEE

was reported that a long training time was needed because it required a large number of computational resources.

The model named SGUNet was developed by Pan et al. [6] and is a modified UNet architecture with a Semantic Guidance Module (SGM). The network consists of an encoder-decoder structure with five convolutional stages using average pooling for downsampling and leaky ReLU for activation. SGMs are integrated into all decoding layers to perform segmentation by generating single-channel semantic feature maps that guide the upsampling process. SGUNet was tested on the Thyroid Digital Image Database (TDID), a dataset containing 400 B-mode ultrasound images. When SGUNet's test results are examined, it achieves a value of 72.9% in Dice Score and 60.4% in IoU, outperforming traditional U-Net by 2.0% in Dice Score and 2.4% in IoU.

III. Materials and Methods

A. Dataset

The dataset named TN3K [4] used in this study consists of 3493 thyroid ultrasound images and includes ground truth masks for these images. These images were collected with ethical permission from Zhujiang Hospital and South Medical University and are provided as open source. In addition, three different ultrasound devices were used to collect data: GE Logiq E9, ARIETTA 850, RESONA 70B.

B. Methods

The model called U-Net, which is based on convolutional neural networks and designed for biomedical image segmentation, contains two paths with a symmetric U-shaped structure. The first path is called encoder and consists of repeated convolution followed by rectified linear unit (ReLU) and max pooling layers. The purpose of the encoder path is to capture context and reduce the spatial dimension. The path called decoder, on the other hand, enlarges the spatial resolution by performing transposed convolution. With this structure, well-localized segmentation images can be achieved [7].

The UNet++ model, which was improved by adding nested and dense skip connections to the UNet model, can capture multi-level features between the encoder and decoder. With these changes, it is seen that UNet++ provides more precise segmentation compared to U-Net [8].

UNet3+ is a model developed to further improve the segmentation performance of UNet++. It includes triple-skip connections to improve feature learning. It can make feature learning more detailed by adding more output layers. These improvements increase the segmentation accuracy, although they cause increased cost [9].

The VGG16 model added to the encoder section of U-Net models consists of 16 convolution layers with strong feature extraction capability. In the developed hybrid UNet models, the learning rate for training the VGG16 model was selected as 1e-4 and the model was trained for 10 epochs. After the training phase, the weights of the model were recorded for use in the encoder section.

While creating hybrid U-Net models, VGG16 model previously trained with thyroid ultrasound images was added as an encoder. In addition, attention blocks were added to the decoder path. L1 and L2 regularization was used to reduce the sensitivity of hybrid models to complexity. In Hybrid U-Net++, a pre-trained VGG16 model is used as an encoder. A pretrained VGG16 integrated as an encoder in the hybrid UNet3+ architecture performs multi-scale feature fusion by progressively upsampling and combining features from various depths of the encoder to improve the segmentation output. The final output layer generates a segmentation map that makes the model suitable for detailed image segmentation tasks.

IV. Experimental Results

The training and testing stages of the models were performed with K-fold cross-validation. The K value was selected as 5 and the dataset was divided into 5 equal parts in each layer. A different part was used for testing in each layer and the remaining 4 parts were used for training. The models were trained in 30 epochs with a 16 "batch" size and 0.001 learning rate. The binary cross-entropy, which measures the difference between the real labels and the predicted probabilities, was collected with the dice loss, which measures the overlap between the predicted and real segmentation masks, and the combined loss used in training the models was obtained. IoU, Dice

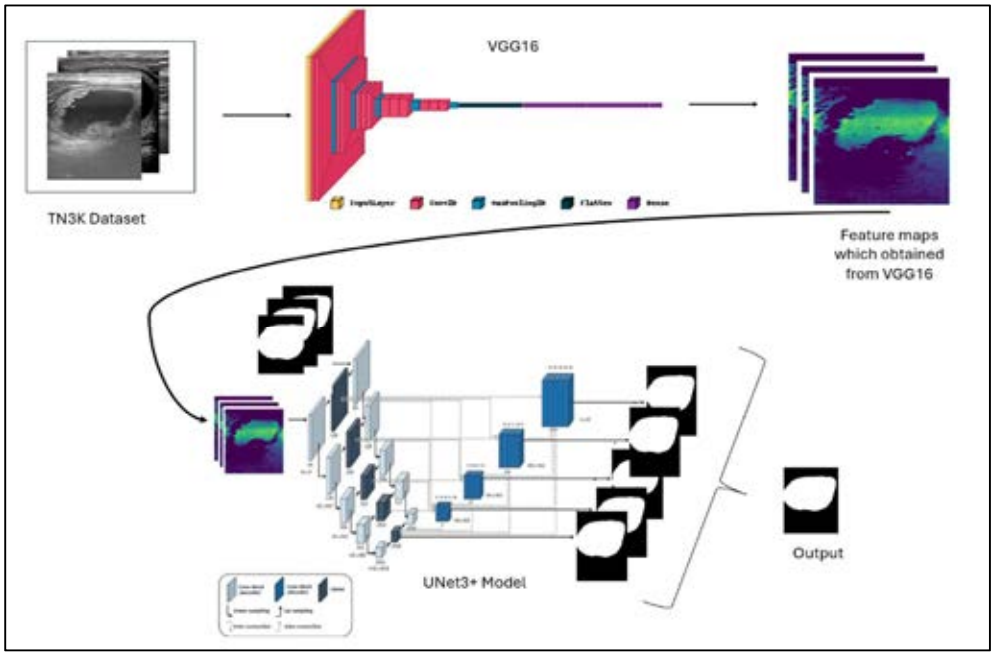


Figure 1. Graphical summary of the Hybrid UNet3+ model created in the study

coefficient score, precision, and recall were used to evaluate the performance of the models, and the formulas for these metrics are given in equations 1, 2, 3, and 4, respectively. The value indicated by A in the formulas is the pixel set in the predicted mask and the pixel set in the ground truth mask indicated by B. TP represents true positive, FP false positive, and FN false negative values.

- **IoU (Intersection over Union):** It measures the overlap between predicted and true masks based on their combination [11].

$$IoU = \frac{|AUB|}{|A \cap B|} \quad (1)$$

- **Dice Score:** It measures the overlap between predicted and true masks, with emphasis on the intersection [12].

$$Dice = \frac{|A| + B|}{2 * |A \cap B|} \quad (2)$$

- **Precision:** Measures the proportion of correctly predicted positive pixels out of all pixels that were predicted as positive [13].

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

- **Recall:** Measures the proportion of correctly predicted positive pixels out of all actual positive pixels in the ground truth [13].

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

The models were trained and tested with 5-fold cross-validation, the values for each fold were collected and the average values of the 5 folds were taken [14]. Accordingly, the IoU, Dice Score, precision, and recall average values of the 6 models are given in Table 1.

TABLE I. MODELS PERFORMANCES

Models	IoU	Dice Score	Precision	Recall	Loss
UNet	50.431	62.602	63.327	72.594	0.576
UNet++	61.321	72.944	73.873	79.377	0.429
UNet3+	63.560	74.796	75.540	81.086	0.411
Hybrid UNet	58.827	70.909	76.394	73.634	3.542
Hybrid UNet++	64.729	75.546	78.231	79.792	0.408
Hybrid UNet3+	**65.404**	**76.271**	**79.222**	**80.071**	**0.398**

When looking at the IoU value, UNet gave a value of 50.431, while Hybrid UNet showed great improvement and reached a value of 58.827. While 61.321 IoU value was obtained with the UNet++ model, 64.729 value was obtained with the Hybrid UNet++ model. The model that reached the highest IoU value among U-Net models was the UNet3+ model with 63.560. However, the hybrid UNet3+ model exceeded this value and managed to reach 65.404 IoU value.

When looking at the Dice Score, the U-Net model achieved a score of 62,202, while the Hybrid U-Net model significantly exceeded this score and achieved a Dic Score of 70,909. While the Dice Score of the UNet++ model was 72,944, the Hyrid UNet++ model increased this score to 75.5461. The UNet3+ model, which achieved a Dice Score of 74,796, achieved the highest segmentation performance with the Hybrid UNet3+ model, with the highest Dice Score of 76,271.

In terms of Sensitivity, which measures the accuracy of positive predictions, the regular U-Net model scored 63.327, while the Hybrid U-Net model improved to 76.394. The UNet++ model had a Sensitivity of 73.873, and the Hybrid UNet++ model improved it to 78.231. The UNet3+ model had a Sensitivity of 75.540, but the Hybrid UNet3+ model improved it to 79.222.

When we look at the Recall, which measures the model's ability to detect all true positives, the standard U-Net model achieved 72,594, with the Hybrid U-Net model showing a slight improvement with a Recall of 73,634. The UNet++ model recorded a Recall of 79,377, with the Hybrid UNet++ model slightly improving it to 79,792. The UNet3+ model had the highest Recall of 81,086 among all the regular models, while the Hybrid UNet3+ model achieved a slightly lower Recall of 80,071.

The values of the Loss metric, which shows the error in the model's predictions, are given in Table 1 again. The Normal U-Net model has a Loss of 0.576, but the Hybrid U-Net model shows a much higher Loss of 3.542, which is an unexpected and significant increase. On the other hand, the UNet++ model has a Loss of 0.429, and the Hybrid UNet++ model slightly reduces it to 0.408. The Loss of the UNet3+ model is 0.411, and the Hybrid UNet3+ model has the lowest Loss of 0.398.

V. DISCUSSION

When the results were evaluated, Hybrid U-Net models generally provided improvements over their normal counterparts, especially in IoU, Dice Score, and Precision. However, the increased Loss in the Hybrid U-Net model indicated potential difficulties in model optimization, and these difficulties were successfully reduced in the more advanced Hybrid UNet++ and Hybrid UNet3+ models. Additional improvements were made in the hybrid models, such as more complex encoder structures and multi-scale feature extraction capabilities of hybrid models. Thanks to these improvements, thyroid nodule segmentation with detailed features was improved. In short, the Hybrid UNet3+ model can be used in thyroid nodule diagnosis by providing positive results in the thyroid segmentation task.

Additionally, the graph in Figure 2 shows the validation of the Dice score for the hybrid UNet3+ model over 30 epochs in

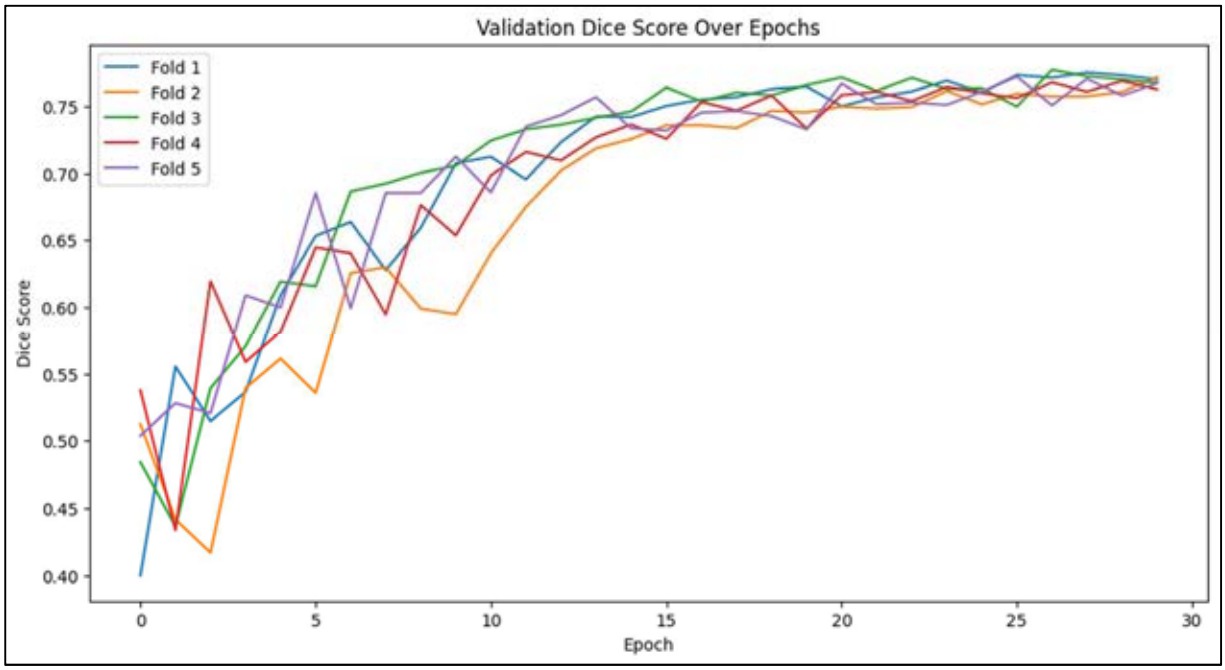


Figure 2. Hybrid UNet3+ models validation Dice Score graph over folds

five different folds. Dice score increases rapidly in early epochs, indicating that the model quickly learns to segment images effectively. As epochs progress, the Dice score stabilizes and reaches a consistent value of around 0.75 to 0.78, indicating that the model achieves good segmentation performance. Overall, the model shows strong and consistent performance in all folds.

Thyroid nodule segmentation is an important task in medical imaging, especially in the diagnosis and treatment of thyroid-related diseases. Accurate segmentation of thyroid nodules from ultrasound images can reduce the workload of doctors and can be used for rapid diagnosis of patients. In future studies, the malignancy risk of nodules can be evaluated, and accurate treatment planning can be performed. In addition, nodule size, shape, and volume, which are basic parameters that will facilitate the distinction between malignant and benign, can be precisely measured automatically with deep learning.

REFERENCES

[1] Kara, İ., Yıldız, M. G., & Orhan, İ. (2020). Tiroid Nodülü. KSU Medical Journal, 15(3), 94-99.

[2] Şengöz, T., Çubuk, R., Kaya, H., & Arıbal, E. (2009). Ultrasonography Guided Fine Needle Aspiration Biopsies On Thyroid Nodules. Düzce Tıp Fakültesi Dergisi, 11(2), 26-32.

[3] Welker, M. J., & Orlov, D. (2003). Thyroid nodules. American Family Physician, 67(3), 559-567.

[4] 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.

[5] Ozcan, A., Tosun, Ö., Donmez, E., & Sanwal, M. (2024). Enhanced-TransUNet for ultrasound segmentation of thyroid nodules. Biomedical Signal Processing and Control, 95, 106472.

[6] Pan, H., Zhou, Q., & Latecki, L. J. (2021, April). Sgunet: Semantically guided unet for thyroid nodule segmentation. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 630-634). IEEE.

[7] Krithika Alias AnbuDevi, M., & Suganthi, K. (2022). Review of semantic segmentation of medical images using modified architectures of UNET. *Diagnostics*, *12*(12), 3064.

[8] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4* (pp. 3-11). Springer International Publishing.

[9] Xu, Y., Hou, S., Wang, X., Li, D., & Lu, L. (2023). A medical image segmentation method based on an improved UNet 3+ network. *Diagnostics*, *13*(3), 576.

[10] Chhabra, M., & Kumar, R. (2022). An advanced VGG16 architecture-based deep learning model to detect pneumonia from medical images. In *Emergent Converging Technologies and Biomedical Systems: Select Proceedings of ETBS 2021* (pp. 457-471). Singapore: Springer Singapore.

[11] Cheng, B., Girshick, R., Dollár, P., Berg, A. C., & Kirillov, A. (2021). Boundary IoU: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15334-15342).

[12] Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., & Blaschko, M. B. (2019). Optimizing the dice score and Jaccard index for medical image segmentation: Theory and practice. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22* (pp. 92-100). Springer International Publishing.

[13] Erdaş, Ç. B. (2024). Computer-aided colorectal cancer diagnosis: AI-driven image segmentation and classification. *PeerJ Computer Science*, *10*, e2071.

[14] ERDEMİR, E., & ERDAŞ, Ç. B. (2023). Polikistik Over Sendromu Tanısında Derin Öğrenme Tekniklerinin Değerlendirilmesi Evaluation of Deep Learning Techniques in the Diagnosis of Polycystic Ovary Syndrome.