AIMS Medical Science

AIMS Medical Science, 12(2): 124–144.
DOI: 10.3934/medsci.2025009
Received: 27 September 2024
Revised: 10 March 2025
Accepted: 20 March 2025
Published: 03 April 2025

http://www.aimspress.com/journal/medicalScience

Research article

Thyroid nodule segmentation in ultrasound images using U-Net with ResNet encoder: achieving state-of-the-art performance on all public datasets

Antonin Prochazka* and Jan Zeman

Institute of Biophysics and Informatics, 1st Faculty of Medicine, Charles University, Prague, Salmovska 1, Praha 2, 120 00, Czech Republic

* **Correspondence:** Email: antonin.prochazka@lf1.cuni.cz.

Abstract: Ultrasound imaging plays a vital role in evaluating thyroid nodules, aiding in the assessment of malignancy risk, monitoring size progression, and serving as a guiding tool for thyroid nodule biopsies. Computer-aided diagnosis (CAD) systems have emerged to assist in diagnosing thyroid lesions, reducing unnecessary biopsies, and contributing to the overall improvement of diagnostic accuracy. The segmentation process plays a crucial role in CAD systems because it marks the region of interest. If segmentation were sufficiently accurate, then it would improve the entire diagnostic process and bring CAD systems closer to routine clinical practice. As far as we know, there are currently only three publicly available datasets of ultrasound images of the thyroid gland that can be used for the purpose of thyroid nodules segmentation. The Thyroid Digital Image Database (TDID) is a long-standing benchmark dataset but faces limitations due to the data ambiguities. The TN3K dataset is more robust than TDID, and the Thyroid Ultrasound Cine-clip dataset offers recent alternatives. In this paper, we implemented a deep learning segmentation model based on UNet with a ResNet encoder. We trained this model on all available data and evaluated it on the TN3K test set. The achieved results for the Dice score, IoU score, accuracy, precision, and recall were 84.24%, 75.48%, 97.24%, 82.75%, and 88.98%, respectively. These results represent the most advanced state-of-the-art scores compared to previously published studies and demonstrate that UNet with a ResNet encoder has the capability to accurately segment thyroid nodules in ultrasound images.

Keywords: deep learning; segmentation; thyroid nodules; ultrasound; ResNet; UNet

1. Introduction

Ultrasound (US) imaging, typically ranging from 7.5 to 15 MHz, holds significant importance in the evaluation of thyroid nodules. The superficial nature of the thyroid gland makes it highly amenable to US imaging, thus providing us with detailed and clear images with very good spatial resolution. US imaging assumes a critical role in evaluating the malignity risk of nodules in the thyroid gland, monitoring their progression, and serving as a control imaging modality for nodule biopsies.

Thyroid nodules are abnormal growths in the thyroid gland. Most thyroid nodules are benign, though some of thyroid nodules may be cancerous, thus necessitating further evaluation and treatment. Thyroid cancer has seen a significant increase in diagnosis rates worldwide, ranking seventh among the most common cancers in women and fifteenth in men [1,2]. Due to this growing incidence, there is a need for effective diagnostic tools, and B-mode ultrasound imaging has emerged as the first choice. Thyroid nodules can be categorized as solid, cystic, or a combination of both solid and cystic components. Epidemiological studies indicate that palpable thyroid nodules are present in approximately 5% in women and 1% in men [3]. However, when ultrasound examinations are conducted, the incidental detection rate of thyroid nodules significantly increases to a range of 19% to 67% [4,5].

There has been significant research focused on evaluating and ranking the ultrasound risk features used by endocrinologists to predict the malignant potential of thyroid nodules. Several studies have investigated these risk features, and two meta-analyses were conducted by Remonti et al. [6] and Brito et al. [7] synthesized the findings from a large number of studies. Remonti et al. included over 12,500 nodules from 54 studies in their meta-analysis, while Brito et al. analyzed a total of 18,288 nodules. The results from these meta-analyses indicate that certain US features are associated with an increased risk of malignancy. These features include calcifications, a taller-than-wide shape, irregular margins, absence of elasticity, hypoechogenicity (reduced echogenicity), increased blood flow, absence of a halo, and/or larger nodule size. The research conducted on thyroid nodules and their sonographic features have contributed to the development of the Thyroid Imaging, Reporting, and Data System (TI-RADS) by the American College of Radiology [8].

In addition to research focused on the visual inspection of ultrasound images by endocrinologists, computer-aided diagnosis (CAD) systems have emerged to assist in diagnosing thyroid lesions and reducing unnecessary biopsies. These CAD systems [9–14] utilize US images to facilitate the accurate and efficient classification of benign and malignant thyroid nodules. Typically, CAD systems for thyroid nodules consist of two main components: segmentation and classification. The segmentation part of the system is responsible for accurately identifying and delineating the boundaries of the nodule within the ultrasound image of the thyroid gland. Once the nodule is successfully segmented, the classification component analyzes the extracted features and provides an estimation of the nodule's malignant potential. The segmentation process plays a crucial role in CAD systems as it enables the system to specifically focus on the nodule region of interest. This step helps in excluding irrelevant structures and background noise from the analysis, thus allowing for more accurate classifications. By accurately delineating the nodule boundaries, the CAD system can concentrate on the specific characteristics and features of the nodule that are indicative of its malignancy.

Although conventional ultrasound provides valuable insights into thyroid nodules, it remains a two-dimensional (2D) imaging modality, which presents limitations in tracking nodule changes over time. Three-dimensional (3D) US imaging has the potential to improve the diagnostic accuracy by

offering volumetric information, thus allowing clinicians to more effectively assess nodule growth, changes in structure, and spatial relationships. Tracking nodule volumes over time is essential in clinical decision-making, as volume changes may indicate malignant transformations or nodule regression. Future advancements in segmentation models should explore their applicability in 3D US imaging to enhance the longitudinal assessment and ensure a robust diagnostic performance.

1.1. Segmentation

Before the emergence of deep learning, US image segmentation for thyroid nodules primarily relied on conventional image processing techniques. These techniques can be broadly categorized into three groups: contour and shape-based methods, region-based methods, and traditional machine learning methods. While these conventional techniques made valuable contributions to thyroid nodule segmentation, deep learning-based approaches has revolutionized thyroid nodule segmentation in US images, thus offering improved accuracy on larger datasets by automatically learning features directly from the data. In the review paper authored by Chen et al. [15], these conventional image processing techniques were extensively examined. However, in our current discussion, we will solely focus on deep learning, as the method proposed in our paper specifically pertains to this approach.

1.2. Deep learning methods in segmentation of thyroid nodules

Deep neural networks, particularly convolutional neural networks (CNNs), have demonstrated remarkable performances in various computer vision tasks including segmentation. Several studies reported using them on the segmentation of thyroid nodules with very promising results. The following overview reports results of image segmentation methods, whether employed individually or as a part of the entire CAD system. Unfortunately, direct comparisons are challenging since, in the majority of cases, the authors employed private datasets. An exception to this trend is the study conducted by Gong et al. [16] and few others [17–19]. We explicitly mention when the authors utilized a public dataset, such as the Thyroid Digital Image Database (TDID) [20].

Ma et al. [21] developed a CNN model with 15 convolutional layers and 2 max pooling layers. The input image size of CNN model was 353 × 353 and the output image size was 44 × 44. Using a dataset of 22,123 US images, the model achieved a mean overlap value of 86.83% on a test set. Another approach, proposed by Ying et al. [22], employed a cascaded convolutional neural network (CCNN). The CCNN consisted of three phases: using a U-Net-based [23] and VGG- based [24] model to extract regions of interest (ROIs) containing thyroid nodules, followed by artificial marks and a fully CNN for segmentation. With 1,000 ultrasound images (800 in training set and 200 the testing), the CCNN achieved a mean overlap value of 87.00% on a test set. Kumar et al. [25] proposed an approach based on dilated convolutional layers that accurately segmented thyroid nodules, cystic components, and normal thyroid gland from ultrasound scans. The algorithm achieved a mean Dice coefficient of 0.76 and demonstrated high detection rates for thyroid nodules and cystic components. Pan et al. [17] proposed a thyroid nodule segmentation approach called SGUNet, that utilized a pixel-wise semantic map to guide low-level features, thus resulting in improved nodule representation. The evaluation on the Thyroid Digital Image Database (TDID) demonstrated SGUNet's superiority over traditional UNet and UNet++ with a 72.9% Dice coefficient. Sun et al. [26] proposed a dual-path CNN with soft shape supervision for the accurate segmentation

of thyroid nodules on US images. The network achieved a high accuracy of 95.81% and a Dice coefficient of 85.33%. Song et al. [27] proposed FDnet, a feature-enhanced dual-branch network. By incorporating a semantic segmentation branch and a feature enhancement mechanism, FDnet improved the proposal scores and reduced the false positives segmentation. The Boundary Attention Transformer Network (BTNet), introduced by Li et al. [18], integrated a segmentation network with a boundary attention mechanism, thus combining the advantages of a convolutional neural network and transformer. The results were an IoU of 0.81 and a Dice score of 0.89 on a private dataset and an IoU 65.4 and a Dice score 75.7 on a TDID dataset. Ataide et al. [28] compared UNet, SUMNET, Attention UNet, and a combination of ResNet and UNet. They found that the combination of ResNet and UNet had the highest Dice score on the private dataset. Gong et al. [16] published TRFE+, a network that utilized the thyroid region prior guided attention for accurate thyroid nodule segmentation in US images with an IoU of 0.71 and a Dice score of 0.83.

1.3. Motivation

The last-mentioned noteworthy paper presented a comprehensive comparison of various architectures, although it overlooked a particularly robust one—the combination of ResNet and UNet. This combined architecture (ResUNet) has established itself as one of the leading state-of-the-art approaches for US image segmentation. Additionally, the authors introduced TN3K, an open-access dataset consisting of high-quality nodule masks, which was strictly divided into the train and test sets. Therefore, it can be used for the direct comparison of segmentation models. Our research aims to implement ResUNet and evaluate its performance in comparison to the results reported by Gong et al. This combination of ResNet and UNet shows the promising results in different segmentation tasks in biomedical area [29–34]. It is important to note the inherent distinctions between US images and images from other biomedical imaging modalities, such as computed tomography (CT) or magnetic resonance (MR) scans. Unlike the latter methods, US image formation relies on different physical quantities, specifically the acoustic impedance of tissues and its differences on tissue boundaries. The process of image formation introduces challenges, as certain boundaries may not be visible (those parallel to the ultrasound waves), or that US creates so called shadows, which are artefacts that can be caused by a strong reflection from a boundary. Additionally, there may be a diffraction of ultrasound waves on boundaries which introduce a significant noise in the US images. Moreover, there are limitations in terms of brightness and contrast in comparison to CT, which uses Hounsfield units to code the overall brightness scale that can be viewed by separate windowed images. Therefore, we assert that optimizing the segmentation of US images needs a more generalized methodology to effectively address the aforementioned challenges. In the context of US image segmentation, the utilization of the ResNet and UNet combination is an area that has yet to be extensively explored. Nevertheless, existing literature suggests that it holds promise as one of the top-performing options. In several studies [28,35,36], it has been found that the combination of ResNet and UNet is the most effective for segmentation in US images. The initial study that compared architectures on US images was conducted by Cai et al. [35]. In this study, the authors assessed the combination of ResNet and UNet, thereby incorporating an attention layer after the ResNet encoder. They conclude that this architecture is more precise than TransUnet [37], which utilizes transformers to encode tokenized image patches from a feature map as the input sequence to extract the global contexts in the neural network. The second paper, authored by Song et al. [36], compared architectures for the segmentation

of US images of the kidney. The study concluded that the combination of ResNet and UNet achieved the second-best dice score, following DeepLabV3+ (which also incorporates the ResNet backbone). The third paper [28] corresponds to the previously mentioned study, where various architectures were used on a private dataset of the thyroid gland. The authors' results indicated that ResUNet achieved the highest Dice score.

2. Materials and methods

2.1. Architecture

The architecture of our model combines UNet with a ResNet34 used as an encoder. This integrated design, referred to as ResUNet, represents a synthesis of ResNet and UNet, strategically devised to leverage the advantages of both architectures. In this design, the traditional UNet's encoder, composed of convolutional layers and max-pooling, is replaced by ResNet34, which uses a series of residual blocks. Each residual block contains two 3×3 convolutions with batch normalization and ReLU, and also adds the block's input to its output via a shortcut connection. This mechanism not only facilitates the training of deeper networks by improving the gradient flow, but also constrains the learned modifications to be minor, thus preserving critical information. Following the encoding phase, the decoder mirrors the UNet structure. It upsamples the feature maps using 2 × 2 convolutions and employs skip connections to concatenate the corresponding encoder features. This blend of high-level contextual information with fine-grained spatial details allows for precise localization, culminating in a segmentation head that applies a final convolution and sigmoid activation to produce the output map. The use of residual connections in the encoder mitigates vanishing gradients and allows for deeper, more robust network designs. Skip connections from the encoder to the decoder facilitate the combination of contextual and detailed spatial features, which can potentially enhance the segmentation accuracy. By combining ResNet's powerful feature extraction with UNet's efficient localization strategy, ResUNet offers a compelling balance that is well-suited for complex segmentation tasks.

However, ResUNet has some limitations, particularly in terms of data sensitivity and hyperparameter sensitivity. The deeper architecture and increased parameter count can lead to overfitting if the model is not trained on sufficiently large and diverse datasets, thus making effective data augmentation and hyperparameter optimization crucial to ensure that ResUNet reaches its full potential in image segmentation tasks. The architecture of our ResUNet is show in the Figure 1.

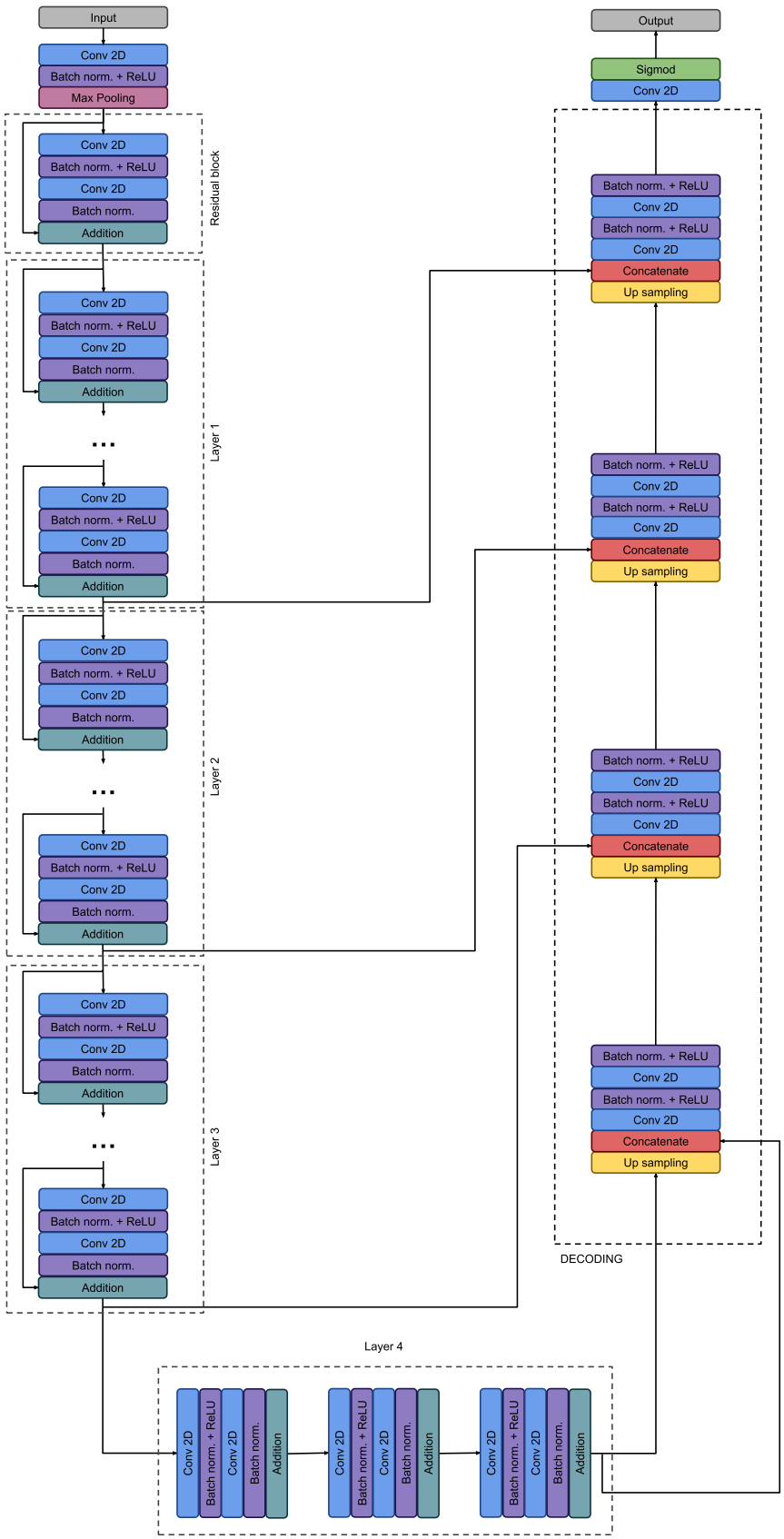

Figure 1. The figure illustrates the architecture of our ResUNet. The ResUNet is a deep learning model that combines the concepts of both ResNet and UNet.

2.2. Dataset

As far as we know, all publicly available datasets of images of thyroid nodules are used in this study. They are three in total: the TN3K dataset, the Thyroid Digital image Database (TDID) dataset, and the Thyroid Ultrasound Cine-clip dataset.

2.2.1. TN3K dataset

This dataset[1] was published by Gong et al. [16] and is divided into a training set and a test set based on the criterion that images from the same patient only appear in one specific subset. The training set consists of 2,879 images, while the test set contains 614 images. We adhered to this data partitioning and utilized the TN3K test set to evaluate the performance of our models. This enabled us to make direct comparisons between our results and those published by Gong et al.

2.2.2. Thyroid Digital Image Database (TDID) dataset

This dataset[2] was published by Pedraza et al. [20] and it was used in several studies for segmentation [17–19] and classification [38,39]. This dataset is comprised of images along with corresponding XML files containing coordinates. However, during the conversion process from coordinates to binary masks, we encountered several issues. Specifically, not all the coordinates were present in the XML files, and some problems arose with the masks themselves (e.g., some of them are outside of ultrasound image area). As a result, it was necessary to either adjust the coordinates of certain masks or remove certain images entirely. A comprehensive list of deleted or modified images is provided in Appendix A.

2.2.3. Thyroid Ultrasound Cine-clip dataset

This dataset[3] is comprised of 167 patients who have been biopsy-confirmed to have thyroid nodules (n = 192) at Stanford. The dataset includes US cine-clip images, segmentations annotated by radiologists, patient demographics, lesion size and location, TI-RADS descriptors, and histopathological diagnoses. The total number of cine-clip frames is 17,412. The dataset is comprised of thyroid nodule cine-clip sequences with frame counts ranging from a minimum of 11 frames to a maximum of 442 frames per nodule. The dataset distribution is shown in Figure 2. Given that the dataset is comprised of multiple images (frames) per thyroid nodule, our objective is to investigate how the evaluation metrics are influenced by the number of frames (images) assigned to each thyroid nodule. Since the number of frames is not constant for each nodule, we employed the following sampling strategy. When extracting n frames from the dataset, we address nodules with frame numbers below n by incorporating all available frames. Conversely, for nodules with occurrences in the dataset where the frame counts exceed n, we opt for a random selection of n frames corresponding to the specific nodule.

[1] Available here: https://github.com/haifangong/TRFE-Net-for-thyroid-nodule-segmentation.

[2] Available here: http://cimalab.unal.edu.co/applications/thyroid/.

[3] Available here: https://stanfordaimi.azurewebsites.net/datasets/a72f2b02-7b53-4c5d-963c-d7253220bfd5.

Image examples from the 3 datasets and the corresponding ground truth masks are shown in Figure 3.

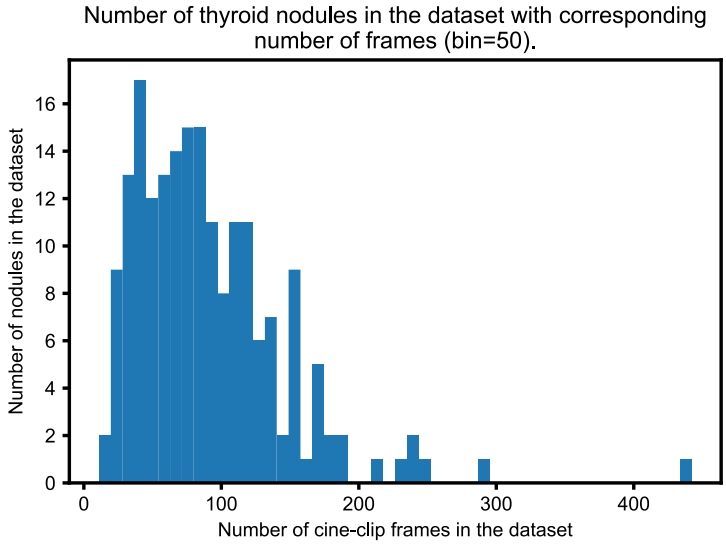

Figure 2. A histogram showing the distribution of the cine-clip frames in the Thyroid Ultrasound Cine-clip dataset.

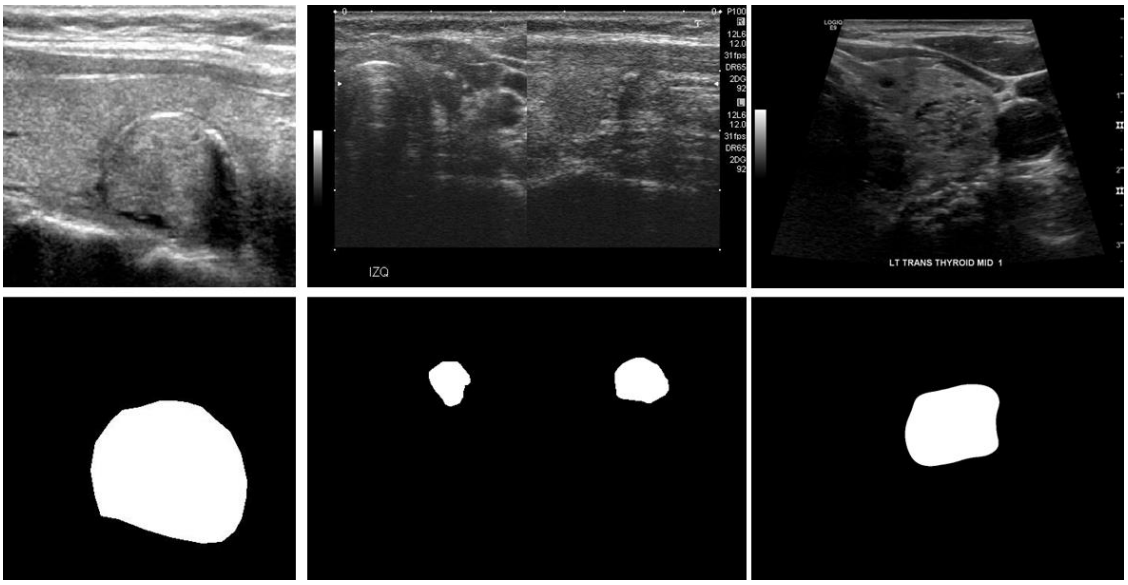

Figure 3. This figure illustrates examples of training images. The first image is from the TN3K dataset, the second image is from the TDID dataset, and the third image is sourced from the Thyroid Ultrasound Cine-clip dataset (Table 1). These images vary in size and represent different views of the thyroid gland. The first image captures the thyroid gland in the longitudinal view, while the third image depicts an axial view. The second image is a combination of both views in split-screen mode, with the left image showing the axial view and the right image showing the longitudinal view. Additionally, the third image was acquired using a sector probe.

Table 1. Thyroid image datasets used in this study.

Dataset	Train	Test	Ultrasound imaging device
TDID	464	-	TOSHIBA Nemio 30, TOSHIBA Nemio MX
TN3K	2879	614	GE Logiq E9, ARIETTA 850, RESONA 70B
Thyroid US Cine-clip	192–17,412	-	GE Logiq E9

2.3. Data augmentation

Data augmentation is employed in image segmentation to increase the diversity and quantity of the training data. By applying transformations such as resizing, flipping, rotation, and contrast adjustments to the original images, the augmented dataset provides the model with more varied samples to learn from. This process helps improve the segmentation model's robustness, generalization ability, and performance on unseen data. We implement augmentations using Albumentations library [40] on each image and its mask in the training data (after division to training and validation set), including random adjustments of brightness and contrast (up to 35%, with probability of 50%), shifts, scales, and rotations (limited by 35%, 35%, and 35 degrees, respectively) with a probability of 50% and a horizontal flip with a probability of 50%. Each image and its mask undergo augmentations, giving it a 50% chance of being modified by one of the augmentation functions. Consequently, there is a 12.5% probability that an image will remain unaltered. If shifts, scales, and rotations affect the integrity of the original reference region, then the same adjustments are applied to the corresponding masks. This might result in the nodule being positioned at the edge of the image, which is a scenario that is not uncommon in original datasets where nodules can naturally appear at the image edge. It's important to note that we do not generate new images through augmentations: we solely modify the existing ones. We avoid applying these modifications to the validation and test data.

2.4. Evaluations metrics

To quantitatively assess the segmentation performance of our proposed method, we have chosen several metrics as follows (where TP, FP, TN, FN indicate true positive, false positive, true negative, and false negative, respectively):

- Intersection Over Union (IoU): This measures the ratio of the overlapping area between the predicted and ground truth segmentation masks to the total area encompassed by both masks. It is calculated as follows: TP/(FP + FN).
- Dice Coefficient: The Dice score quantifies the similarity between the predicted and ground truth segmentation masks. It is defined as follows: 2 * TP/(FP + FN + 2 * TP).
- Accuracy: Accuracy evaluates the overall correctness of the segmentation by considering both true positive and true negative predictions, given as: (TN + TP)/(TN + TP + FN + FP).
- Precision: Precision measures the accuracy of positive predictions made by the model. It is calculated as follows: TP/(TP + FP).
- Recall (also known as Sensitivity): Recall measures the ability of the model to correctly identify positive instances from the ground truth. It is defined as follows: TP/(TP + FN).

While human interpretation remains the gold standard, intraobserver and interobserver variability among radiologists is a well-known issue. Automated systems with a high Dice coefficient can help reduce observer-dependent inconsistencies and standardize diagnoses, particularly in large-scale screening applications. To compare the studies that assess interobserver variability—typically using the percentage difference and Bland-Altman limits of agreement (LOA), which measure the percentage agreement between two observers regarding two dimensions of a nodule—we introduce a novel approximation between the percentage difference and the Dice score for elliptic shapes. Let's consider a reference ellipse with axes a and b. Its area is as follows:

$$S_1 = \pi \cdot \frac{a \cdot b}{4} \tag{1}$$

If a second observer measures an ellipse with axes that are scaled versions of the reference, as defined by the scaling factors αa and βb (where α and β represent the percentage of the reference measurements), then

$$S_2 = \pi \cdot \frac{(\alpha a) \cdot (\beta b)}{4} = \alpha\beta S_1 \tag{2}$$

Assuming that both ellipses are concentric (i.e., the second ellipse is entirely contained within the reference ellipse), the area of their intersection is simply S_2.

Then, the Dice coefficient (DSC) is calculated as follows:

$$DSC = \frac{2S_{intersection}}{S_1 + S_2} = \frac{2S_2}{S_1 + S_2} = \frac{2\alpha\beta S_1}{S_1 + \alpha\beta S_1} \tag{3}$$

By canceling out S_1 from the numerator and denominator, we obtain the following general formula:

$$DSC = \frac{2\alpha\beta}{1 + \alpha\beta} \tag{4}$$

This should be considered as an approximation because thyroid nodules do not necessarily have an elliptic shape, and in some cases, the two shapes delineated by observers may not be in the same location, meaning these nodules cannot be considered concentric.

2.5. Setup

We used the Google Colab platform for training and testing, thereby utilizing the NVIDIA Tesla V100 GPUs with a memory capacity of 16 GB. The implementation framework chosen was PyTorch 2.0.1, integrated with CUDA 12.0. Pre-trained encoders from the ImageNet dataset were employed to initialize the model weights. Optimization of the models was accomplished using the Adam algorithm, and a total of 100 epochs were executed during the training phase. We used the Cosine Annealing Learning Rate strategy with a maximal learning rate 0.0005. The Cosine Annealing Learning Rate strategy is defined as follows [41]:

$$lr = lr_{min} + \frac{1}{2}(lr_{max} - lr_{min})\left(1 + \cos\left(\frac{T_{cur}}{T_{max}}\pi\right)\right) \tag{5}$$

where lr is learning rate to be used in the current epoch, lr_{min} is the minimum learning rate, lr_{max} is the maximum learning rate, T_{cur} is the current epoch, and T_{max} is the total number of epochs in the training process. For each epoch, a batch size of 64 was used, and all input images were resized to dimensions of 224 × 224 pixels. For model prediction, all images were simply resized to 224 × 224 without any additional augmentation. The data was divided into training and validation sets in an 80:20 ratio when the training set was only the TN3K dataset; alternatively, the data was in a 70:30 ratio when we utilized other datasets for training. This adjustment aimed to ensure that the validation set encompassed a more diverse range of images, thereby preventing the risk of overfitting.

3. Results

We trained ResUNet solely on the TN3K train set and evaluated its performance on the TN3K test set. Our findings demonstrated superior scores compared to the results published by Gong et al. Specifically, our mean Dice score, mean IoU, and mean accuracy reached 83.77%, 75.09%, and 97.18%, respectively. Moreover, the proposed method involves the direct segmentation of thyroid nodules from thyroid images. In contrast, the TEFE+ method relies on the processing of both thyroid gland and thyroid nodule images. Both types of images pass through a shared encoder; then, they diverge into separate decoders for thyroid nodule segmentation and thyroid gland segmentation. This implies that thyroid gland images are required for training the TEFE+ architecture. In the proposed method, we only use thyroid nodule images. The results are shown in Table 2.

Table 2. Performance of our model on TN3K test set with a comparison to results of other models published by Gong et al. [16]. Results represent percentage score ± standard deviation.

	Train set	Accuracy	IoU	Dice
UNet [23]	TN3K train	96.46 ± 0.11	65.99 ± 0.66	79.51 ± 1.31
SGUNet [17]	TN3K train	96.54 ± 0.09	66.05 ± 0.43	79.55 ± 0.86
TRFE [42]	TN3K train	96.71 ± 0.07	68.33 ± 0.68	81.19 ± 1.35
FCN [43]	TN3K train	96.92 ± 0.04	68.18 ± 0.25	81.08 ± 0.50
SegNet [44]	TN3K train	96.72 ± 0.12	66.54 ± 0.85	79.91 ± 1.69
Deeplabv3+ [45]	TN3K train	97.19 ± 0.05	70.60 ± 0.49	82.77 ± 0.98
CPFNet [46]	TN3K train	97.17 ± 0.06	70.50 ± 0.39	82.70 ± 0.78
TransUNet [37]	TN3K train	96.86 ± 0.05	69.26 ± 0.55	81.84 ± 1.09
TRFE+ [16]	TN3K train	97.04 ± 0.10	71.38 ± 0.43	83.30 ± 0.26
ResUNet (our)	TN3K train	97.18 ± 0.03	75.09 ± 0.22	83.77 ± 0.20

Moreover, we explored different augmentation techniques and hyperparameter settings. When no data augmentation was applied, the mean IoU score was 69.71% ± 23.73% and the mean Dice score was 79.46% ± 22.60%. When only horizontal flip was used, the mean IoU score was 70.53% ± 23.21% and the mean Dice score was 80.51% ± 21.23%. When the Cosine Annealing Learning Rate was not applied, the mean IoU score was 73.09% ± 22.51% and the mean Dice score was 82.31% ± 21.77%. These results are reported on the TN3K test set. Our findings indicate that appropriate augmentation and hyperparameter tuning can improve the segmentation scores by up to 3% on the Dice score and up to 4.5% IoU in our case.

Additionally, we trained ResUNet solely on the TDID dataset, where we evaluated it using 5-fold cross validation (same technique as was reported in Pan et al. publication [17]), and achieved an IoU of 0.7281 and a Dice score of 0.8329. The comparison of the existing results on the TDID dataset is shown in Table 3. It is essential to emphasize that performing a direct comparison on the TDID dataset presents challenges due to variations in the data cleaning process. Throughout the conversion from coordinates to binary masks, we encountered several issues, as detailed in the Datasets section. The issues in the conversion process, as also reported by Pan et al. [17], resulted in a decreased number of images. A comprehensive list of deleted or modified images is provided in Appendix A.

Table 3. Performance of our model on the TDID dataset. Results represent percentage score ± standard deviation (if reported).

	Train set	IoU	Dice
SGUNet [17]	TDID	60.0	72.9
BTNet [18]	TDID	65.4	75.7
MSAC-Unet [19]	TDID	67.3 ± 1.3	79.2 ± 0.8
ResUNet (our)	TDID	72.8 ± 3.0	83.3 ± 1.8

Table 4. The table presents results obtained by varying the size of the training set. The first row corresponds to the outcomes achieved when training solely on the TN3K dataset, which coincides with the last row of Table 2. In the second row, we combined two datasets, TN3K and TDID, and report the corresponding results. Rows three and onward showcase the results obtained by combining all three datasets, i.e., TN3K, TDID, and the Thyroid Ultrasound Cine-clip dataset.

Train set	Total train images	Dice score	IoU	Accuracy	Precision	Recall
TN3K	2879	0.8378 ± 0.20	0.7509 ± 0.22	0.9718 ± 0.03	0.8230 ± 0.21	0.8914 ± 0.19
TN3K + TDID	3343	0.8324 ± 0.20	0.7410 ± 0.20	0.9705 ± 0.04	0.8087 ± 0.21	0.8894 ± 0.21
TN3K + TDID + Cine-clip (n=1)	3535	0.8424 ± 0.19	0.7548 ± 0.21	0.9724 ± 0.03	0.8275 ± 0.20	0.8898 ± 0.19
TN3K + TDID + Cine-clip (n=2)	3727	0.8401 ± 0.19	0.7509 ± 0.21	0.9714 ± 0.04	0.8285 ± 0.20	0.8875 ± 0.19
TN3K + TDID + Cine-clip (n=4)	4111	0.8367 ± 0.19	0.7466 ± 0.21	0.9703 ± 0.04	0.8144 ± 0.20	0.8968 ± 0.19
TN3K + TDID + Cine-clip (n=8)	4879	0.8377 ± 0.20	0.7496 ± 0.22	0.9708 ± 0.04	0.8257 ± 0.20	0.8811 ± 0.21
TN3K + TDID + Cine-clip (n=16)	6407	0.8339 ± 0.20	0.7456 ± 0.22	0.9708 ± 0.04	0.8238 ± 0.20	0.8808 ± 0.21
TN3K + TDID + Cine-clip (n=32)	10955	0.8378 ± 0.20	0.7517 ± 0.22	0.9710 ± 0.04	0.8308 ± 0.20	0.8837 ± 0.21
TN3K + TDID + Cine-clip (n=64)	14106	0.8321 ± 0.21	0.7454 ± 0.22	0.9707 ± 0.03	0.8266 ± 0.20	0.8761 ± 0.22
TN3K + TDID + Cine-clip all	20755	0.8321 ± 0.20	0.7447 ± 0.22	0.9694 ± 0.04	0.8303 ± 0.20	0.8766 ± 0.20

After conducting the initial tests on the TDID and TN3K datasets, we proceeded with further training on other datasets. First, we added the TDID dataset to the training set, and then progressively added the clip frames from the Thyroid Ultrasound Cine-clip dataset. Our objective was to evaluate the model's performance on all known publicly available datasets and assess its performance on the TN3K test dataset to enable a comparison with previously published results. The results are shown in Table 4. For these combined datasets, considering that the Cine-clip dataset contains numerous images (frames) of 192 thyroid nodules, we opted to randomly select a variable number of frames (denoted as "n") from the Cine-clip dataset for each thyroid nodule. In the clarification, we conducted a random sampling process to select one image for each thyroid nodule for the specific case where n = 1. The employed sampling methodology is detailed in the Datasets section.

At the end of training, our model achieved a Dice score of 87.95% on the training set and 86.95% on the validation set, with corresponding IoU scores of 78.59% and 76.98%, respectively. On the test set, it maintained a robust performance with a Dice score of 84.24% and an IoU score of 75.48%. The slight drop of approximately 2.7% in the Dice score and 1.5% in the IoU score from the validation set to the test set indicates a minimal generalization gap, thus suggesting that the model learns robust features that generalize well to unseen data without a significant overfitting.

4. Discussion

The comparison of our implementation of ResUNet on TN3K dataset with previously published works is presented in Table 2. With a different augmentation approach, ResUNet outperforms the other methods. The comparison of ResUNet on the TDID datasets is presented in Table 3. Both results demonstrate that ResNet serves as a powerful encoder for UNet, thereby achieving state-of-the-art results in US image segmentation. Additionally, we explored the possibility of training ResUNet on all publicly available datasets known to us, thus resulting in even better metrics than those achieved solely on the TN3K and TDID datasets. The results are presented in Table 4. From the findings, it is evident that the highest Dice score or IoU were attained by the combined dataset of TN3K and TDID, along with one randomly selected image (n = 1) for each thyroid nodule in the Thyroid Ultrasound Cine-clip dataset.

Despite the application of brightness and contrast augmentation, the final model has incorrectly segmented images with varying the overall intensity. However, it is important to note that some of the images are excessively bright, thus leading to a loss of detail in the bright regions. Another incorrectly segmented images were related to the number of nodules present in the images. When the images contained multiple nodules, the model very often exhibited inaccuracies in its predictions. The model's capability is limited to detecting one or two nodules. The ability to detect up to two nodules can be attributed to the fact that the training data lacked a sufficient number of thyroid glands with 3 or more nodules, thus leading to the limited exposure of the model to such cases. Examples of correctly segmented images are displayed in Figure 4. On the other hand, Figure 5 showcases examples of incorrectly segmented images. In our experiments, we encountered several challenging cases where our segmentation algorithm failed to perform as expected. Specifically, the model struggled with images that exhibited high brightness, multiple nodules, and interference from measurement tools, which resulted in inaccurate segmentation outcomes. We plan to implement optimal augmentation strategies in the future, including controlled variations in the brightness and contrast as well as the simulation of synthetic artifacts, to expand the diversity and representativeness of our training dataset.

We plan to develop methods to detect interfering objects, such as measurement tools and noise, and include similar examples in our training set through data augmentation.

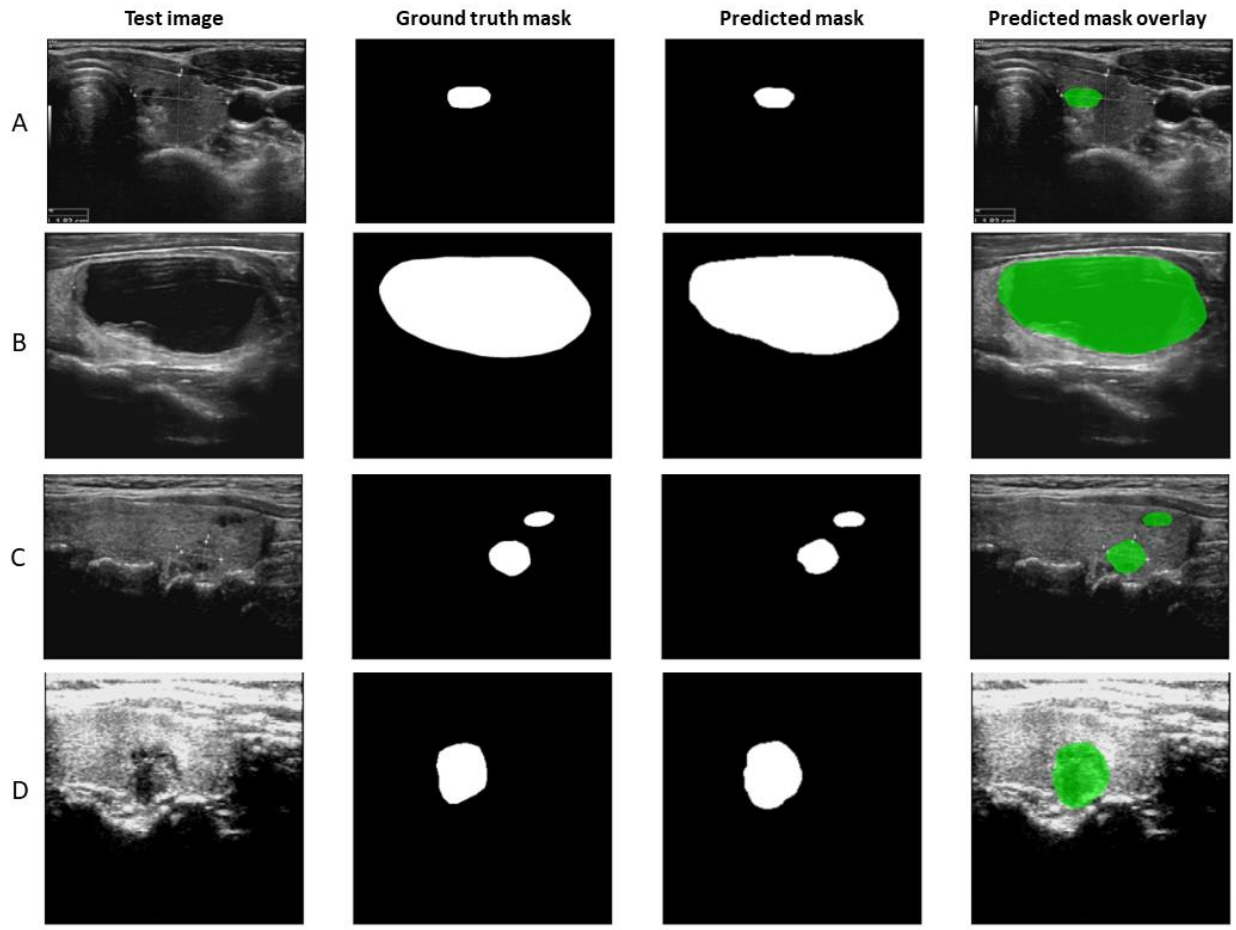

Figure 4. Correctly segmented nodules. The first row (A) displays an image that is being traversed by US measuring tools, posing a common challenge when endocrinologists measure certain features within the image. These measuring tools may identify a nodule itself in some cases, while in other instances, they may indicate different structures, such as the thyroid gland (as illustrated in this case) or the thickness of the isthmus. The segmentation algorithm must distinguish when the measuring tools should be ignored. In the second row (B), a well-segmented cystic nodule is observed, showcasing the excellent performance of our segmentation model on such nodules. In the third row (C), two nodules are accurately segmented, with one being marked by the measuring tools and the other not. In the fourth row (D), an image with significant brightness level is depicted, yet the segmentation model still correctly identified it.

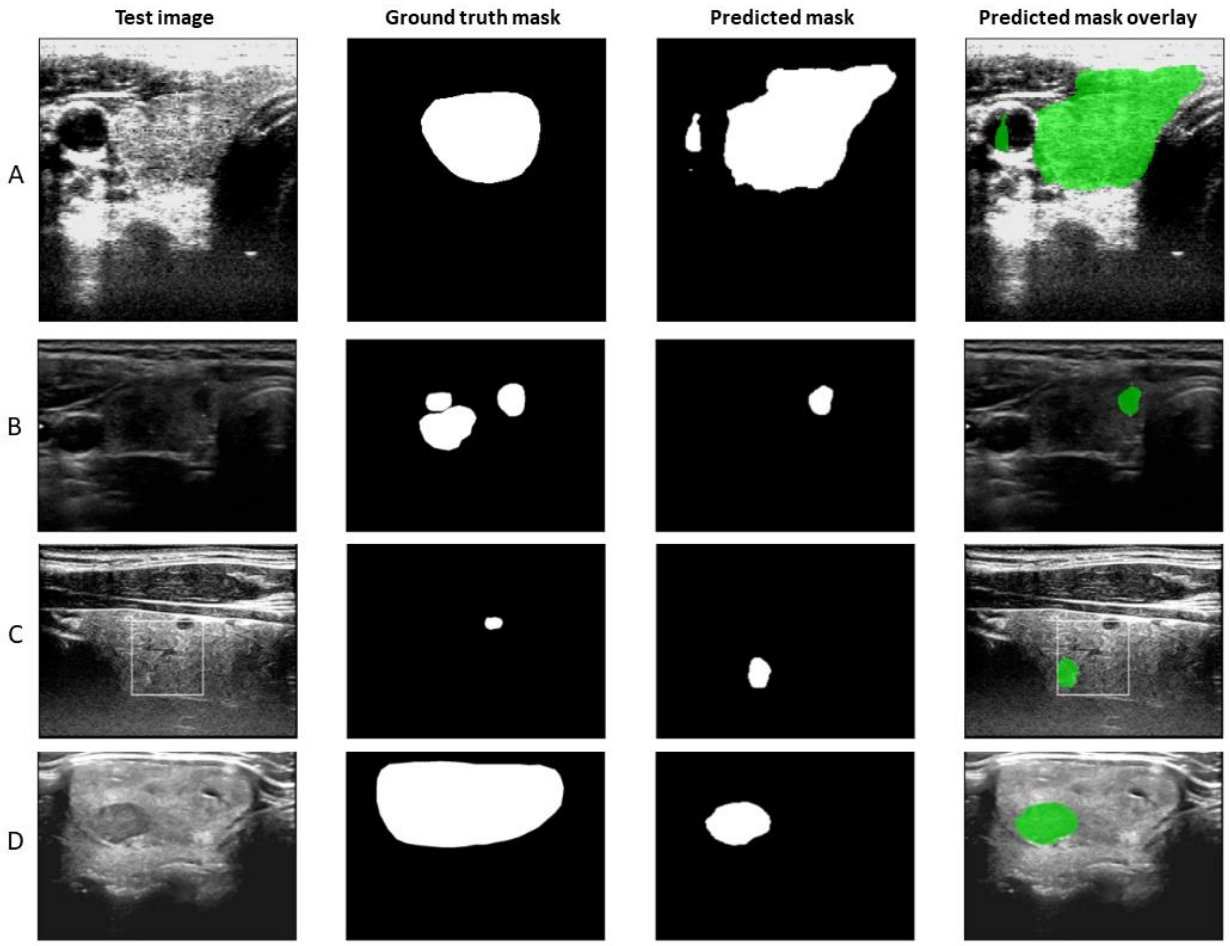

Figure 5. Incorrectly segmented nodules. In the first row (A), there is an image with high brightness, leading to a loss of detail in the bright regions. Despite employing data augmentation during the training of the segmentation model, this particular image remains challenging for the model to accurately segment. In the second row (B), there is an image that contains three nodules, though our segmentation model identified only one. In the third row (C), there is another image that was inaccurately segmented. The measurement tools (the white square) intersect with the nodule, and the nodule appears directly at the edge of the square, which is challenging to segment. In the last row (D), we present an image of a large nodule with local inhomogeneities. The segmentation model successfully identified one of these inhomogeneities, but not the whole nodule.

Moreover, the results demonstrated that using a significant number of almost identical images does not lead to better results when working with a large dataset containing repeated images (Thyroid Ultrasound Cine-clip dataset). We initially hypothesized that augmenting the training set with a larger number of similar but not identical images would result in better performance. Contrary to our expectations, we observed that the trend was opposite. With an increasing number of frames, the Dice score and other metrics were more likely to decrease.

Although CAD systems do not directly challenge human experts, they can ensure a high segmentation performance, thus minimizing the variability and enhancing the reproducibility. In clinical practice, CAD systems are designed to support rather than replace human experts, and higher Dice coefficients indicate a closer alignment between the segmentation outputs and expert annotations. Several studies have assessed the performance of human experts in delineating thyroid nodules. Brauer et al. [47] reported interobserver differences in the mediolateral (21.21%), anteroposterior (20.99%), and craniocaudal (19.89%) diameters. The corresponding interobserver Dice scores, as calculated using the proposed method described in Section 2.4, were approximately 0.77 and 0.78 for the axial and longitudinal planes, respectively. Lee et al. [48] used a Bland-Altman analysis to evaluate the intraobserver and interobserver variability in US measurements of thyroid nodules, thereby reporting percentage differences and the 95% limits of agreement (LOA). When converted to Dice scores using the proposed calculation, the intraobserver Dice scores were identical in both planes: 0.93 for the axial plane and 0.94 for the longitudinal plane. The interobserver Dice scores were 0.92 for the axial plane and 0.91 for the longitudinal plane. Another study [49] focused on the thyroid gland in children. The authors reported that, when taking the mean value of the left and right thyroid lobes, the interobserver differences in lobe measurements (not the nodules) were as follows: a mediolateral difference of 9.7%, an anteroposterior difference of 11.0%, and a craniocaudal difference of 13.7%. The corresponding Dice scores were 0.89 for the axial plane and 0.87 for the longitudinal plane. Our segmentation algorithm achieved a Dice score of 0.84 on a mix of axial and longitudinal images (when compared with expert-annotated ground truth masks in the test set of Gong et al. study [16]). While this result does not fully reach the intraobserver agreement levels, it is within the range of reported interobserver variability among human experts.

5. Conclusions

In this study, we implemented and evaluated a deep learning-based segmentation model, ResUNet, for thyroid nodule segmentation in US images. By leveraging the advantages of both ResNet and UNet architectures, our model achieved state-of-the-art performance, thus surpassing previously published methods. We demonstrated the effectiveness of our approach through extensive experiments using all publicly available thyroid ultrasound datasets, with our model achieving the highest Dice score and IoU metrics on the TN3K test set.

Our findings highlight that training on a diverse dataset significantly improves the segmentation performance on the TN3K test set, particularly when utilizing a combination of TN3K, TDID, and the Thyroid Ultrasound Cine-clip datasets. However, we observed diminishing results when incorporating a large number of near-identical images from cine-clip sequences, thus suggesting that an optimal balance between dataset diversity and redundancy is essential. The diverse and complex nature of thyroid nodules and US image quality highlight the importance of collecting and publishing further datasets. Having access to diverse datasets amplifies a researchers' ability to evaluate and improve their future models, thus enhancing the generalization and robustness to real-world clinical challenges.

In light of this, we acknowledge the novel dataset published by Gong et al. This dataset is originally divided into train and test subsets which is excellent for a direct comparison of studies. We conducted experiments using this dataset, along with other publicly available datasets, and obtained the following mean evaluation metrics (± standard deviations) on the published test set: Dice score, IoU score, accuracy, precision, and recall of 84.24% ± 0.19, 75.48% ± 0.21, 97.24% ± 0.32, 82.75% ±

0.20, and 88.98% ± 0.19, respectively. These results represent the most advanced state-of-the-art scores compared to previously published studies and demonstrate that UNet with the ResNet encoder has the capability to accurately segment thyroid nodules in ultrasound images.

Our results achieved scores comparable to interobserver studies, suggesting that automated thyroid nodule segmentation using deep learning can reach expert-level performance. While CAD systems are not intended to replace human expertise, they can serve as valuable tools for supporting clinical decision-making and improving diagnostic consistency. Our results show that data augmentation has a significant impact, and raises the question of whether it has a greater influence than the architecture used. Considering that the dataset remains small even with all publicly available data on thyroid nodules (and that data augmentation considerably affects the outcomes), we plan to focus our future work on optimizing data augmentation.

Author contributions

This manuscript was collaboratively prepared by both authors, who have contributed to its focus and content. Antonin Prochazka wrote the main part of the manuscript and the Python script. Both authors have reviewed and approved the final version for publication.

Use of AI tools declaration

During the preparation of this work the author(s) used ChatGPT 4 and Grammarly in order to improve readability and language of the text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Acknowledgments

This research used data provided by the Stanford Center for Artificial Intelligence in Medicine and Imaging (AIMI). AIMI curated a publicly available imaging data repository containing clinical imaging and data from Stanford Health Care, the Stanford Children's Hospital, the University Healthcare Alliance and Packard Children's Health Alliance clinics provisioned for research use by the Stanford Medicine Research Data Repository (STARR).

Conflict of interest

The authors declare no conflict of interest.

References

1. Rahbari R, Zhang L, Kebebew E (2010) Thyroid cancer gender disparity. *Future Oncol* 6: 1771–1779. https://doi.org/10.2217/fon.10.127
2. Haugen BR, Alexander EK, Bible KC, et al. (2016) 2015 American Thyroid Association Management Guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association Guidelines task force on thyroid nodules and differentiated thyroid cancer. *Thyroid* 26: 1–133. https://doi.org/10.1089/thy.2015.0020

3. Tunbridge WM, Evered DC, Hall R, et al. (1977) The spectrum of thyroid disease in a community: the Whickham survey. *Clin Endocrinol (Oxf)* 7: 481–493. https://doi.org/10.1111/j.1365-2265.1977.tb01340.x
4. Tan GH, Gharib H (1997) Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. *Ann Intern Med* 126: 226–231. https://doi.org/10.7326/0003-4819-126-3-199702010-00009
5. Guth S, Theune U, Aberle J, et al. (2009) Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. *Eur J Clin Invest* 39: 699–706. https://doi.org/10.1111/j.1365-2362.2009.02162.x
6. Remonti LR, Kramer CK, Leitao CB, et al. (2015) Thyroid ultrasound features and risk of carcinoma: a systematic review and meta-analysis of observational studies. *Thyroid* 25: 538–550. https://doi.org/10.1089/thy.2014.0353
7. Brito JP, Gionfriddo MR, Al Nofal A, et al. (2014) The accuracy of thyroid nodule ultrasound to predict thyroid cancer: systematic review and meta-analysis. *J Clin Endocrinol Metab* 99: 1253–1263. https://doi.org/10.1210/jc.2013-2928
8. Tessler FN, Middleton WD, Grant EG, et al. (2017) ACR Thyroid Imaging, Reporting and Data System (TI-RADS): white paper of the ACR TI-RADS committee. *J Am Coll Radiol* 14: 587–595. https://doi.org/10.1016/j.jacr.2017.01.046
9. Zhu YC, AlZoubi A, Jassim S, et al. (2021) A generic deep learning framework to classify thyroid and breast lesions in ultrasound images. *Ultrasonics* 110: 106300. https://doi.org/10.1016/j.ultras.2020.106300
10. Chi J, Walia E, Babyn P, et al. (2017) Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *J Digit Imaging* 30: 477–486. https://doi.org/10.1007/s10278-017-9997-y
11. Mei X, Dong X, Deyer T, et al. (2017) Thyroid nodule benignty prediction by deep feature extraction. *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, Washington, DC, USA, 241–245. https://doi.org/10.1109/BIBE.2017.00-48
12. Liu T, Xie S, Yu J, et al. (2017) Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 919–923. https://doi.org/109/ICASSP.2017.7952290
13. Prochazka A, Gulati S, Holinka S, et al. (2019) Patch-based classification of thyroid nodules in ultrasound images using direction independent features extracted by two-threshold binary decomposition. *Comput Med Imaging Graph* 71: 9–18. https://doi.org/10.1016/j.compmedimag.2018.10.001
14. Prochazka A, Gulati S, Holinka S, et al. (2019) Classification of thyroid nodules in ultrasound images using direction-independent features extracted by two-threshold binary decomposition. *Technol Cancer Res Treat* 18. https://doi.org/10.1177/1533033819830748
15. Chen J, You H, Li K (2020) A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images. *Comput Methods Programs Biomed* 185: 105329. https://doi.org/10.1016/j.cmpb.2020.105329
16. Gong H, Chen J, Chen G, et al. (2023) Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Comput Biol Med* 155: 106389. https://doi.org/10.1016/j.compbiomed.2022.106389

17. Pan H, Zhou Q, Latecki LJ (2021) SGUNET: Semantic guided UNET for thyroid nodule segmentation, *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, Nice, France, 630–634. https://doi.org/10.1109/ISBI48211.2021.9434051
18. Li C, Du R, Luo Q, et al. (2023) A novel model of thyroid nodule segmentation for ultrasound images. *Ultrasound Med Biol* 49: 489–496. https://doi.org/10.1016/j.ultrasmedbio.2022.09.017
19. Wang R, Zhou H, Fu P, et al. (2023) A multiscale attentional unet model for automatic segmentation in medical ultrasound images. *Ultrason Imaging* 45: 159–174. https://doi.org/10.1177/01617346231169789
20. Pedraza L, Vargas C, Narvaez F, et al. (2015) An open access thyroid ultrasound-image Database. *10th International Symposium on Medical Information Processing and Analysis*. https://doi.org/10.1117/12.2073532
21. Ma J, Wu F, Jiang T, et al. (2017) Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. *Int J Comput Assist Radiol Surg* 12: 1895–1910. https://doi.org/10.1007/s11548-017-1649-7
22. Ying X, Yu Z, Yu R, et al. (2018) Thyroid nodule segmentation in ultrasound images based on cascaded convolutional neural network, In: Cheng, L., Leung, A., Ozawa, S. (eds), *Neural Information Processing*, Springer, Cham, 373–384. https://doi.org/10.1007/978-3-030-04224-0_32
23. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation, In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer, Cham, 9351: 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
24. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. https://doi.org/10.48550/arXiv.1409.1556
25. Kumar V, Webb J, Gregory A, et al. (2020) Automated segmentation of thyroid nodule, gland, and cystic components from ultrasound images using deep learning. *IEEE Access* 8: 63482–63496. https://doi.org/10.1109/ACCESS.2020.2982390
26. Sun J, Li C, Lu Z, et al. (2022) TNSNet: Thyroid nodule segmentation in ultrasound imaging using soft shape supervision. *Comput Meth Prog Bio* 215: 106600. https://doi.org/10.1016/j.cmpb.2021.106600
27. Song R, Zhu C, Zhang L, et al. (2022) Dual-branch network via pseudo-label training for thyroid nodule detection in ultrasound image. *Appl Intell* 52: 11738–11754. https://doi.org/10.1007/s10489-021-02967-2
28. Gomes Ataide E, Agrawal S, Jauhari A, et al. (2021) Comparison of deep learning algorithms for semantic segmentation of ultrasound thyroid nodules. *Curr Dir Biomed Eng* 7: 879–882. https://doi.org/10.1515/cdbme-2021-2224
29. Niu K, Guo Z, Peng X, et al. (2022) P-ResUnet: Segmentation of brain tissue with Purified Residual Unet. *Comput Biol Med* 151: 106294. https://doi.org/10.1016/j.compbiomed.2022.106294
30. Diakogiannis FI, Waldner F, Caccetta P, et al. (2020) ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *Isprs J Photogramm Remote Sens* 162: 94–114. https://doi.org/10.1016/j.isprsjprs.2020.01.013

31. Wang JJ, Gao J, Ren JW, et al. (2021) DFP-ResUNet: Convolutional neural network with a dilated convolutional feature pyramid for multimodal brain tumor segmentation. *Comput Meth Prog Biomed* 208. https://doi.org/10.1016/j.cmpb.2021.106208
32. Morelli R, Clissa L, Amici R, et al. (2021) Automating cell counting in fluorescent microscopy through deep learning with c-ResUnet. *Sci Rep* 11. https://doi.org/10.1038/s41598-021-01929-5
33. Sabir MW, Khan Z, Saad NM, et al. (2022) Segmentation of liver tumor in CT scan using ResU-Net. *Appl Sci* 12. https://doi.org/10.3390/app12178650
34. Wang R, Shen H, Zhou M (2019) Ultrasound nerve segmentation of brachial plexus based on optimized ResU-Net. *2019 IEEE International Conference on Imaging Systems & Techniques (IST)*. https://doi.org/10.1109/IST48021.2019.9010317
35. Cai L, Li Q, Zhang J, et al. (2023) Ultrasound image segmentation based on Transformer and U-Net with joint loss. *PeerJ Comput Sci* 9: e1638. https://doi.org/10.7717/peerj-cs.1638
36. Song SH, Han JH, Kim KS, et al. (2022) Deep-learning segmentation of ultrasound images for automated calculation of the hydronephrosis area to renal parenchyma ratio. *Investig Clin Urol* 63: 455–463. https://doi.org/10.4111/icu.20220085
37. Chen J, Lu Y, Yu Q, et al. (2021) Transunet: Transformers make strong encoders for medical image segmentation. https://doi.org/10.48550/arXiv.2102.04306
38. Chouiha B, Amamra A (2021) Thyroid Nodules Recognition in ultrasound images based on ImageNet top-performing deep convolutional neural networks, In: Senouci, M.R., Boudaren, M.E.Y., Sebbak, F., Mataoui, M. (eds), *Advances in Computing Systems and Applications*, Springer, Cham, 199: 313–322. https://doi.org/10.1007/978-3-030-69418-0_28
39. Gomes Ataide EJ, Ponugoti N, Illanes A, et al. (2020) Thyroid nodule classification for physician decision support using machine learning-evaluated geometric and morphological features. *Sensors* 20. https://doi.org/10.3390/s20216110
40. Buslaev A, Iglovikov VI, Khvedchenya E, et al. (2020) Albumentations: Fast and flexible image augmentations. *Information* 11. https://doi.org/10.3390/info11020125
41. Loshchilov I, Hutter F (2016) Sgdr: Stochastic gradient descent with warm restarts. https://doi.org/10.48550/arXiv.1608.03983
42. Gong H, Chen G, Wang R, et al. (2021) Multi-task learning for thyroid nodule segmentation with thyroid region prior, *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, Nice, France, 257–261, https://doi.org/10.1109/ISBI48211.2021.9434087
43. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell* 39: 640–651. https://doi.org/10.1109/TPAMI.2016.2572683
44. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell* 39: 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615
45. Chen LC, Zhu Y, Papandreou G, et al. (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation, In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) *Computer Vision – ECCV 2018*, Springer, Cham, 11211: 833–851. https://doi.org/10.1007/978-3-030-01234-2_49
46. Feng S, Zhao H, Shi F, et al. (2020) CPFNet: Context pyramid fusion network for medical image segmentation. *IEEE Trans Med Imaging* 39: 3008–3018. https://doi.org/10.1109/TMI.2020.2983721

47. Brauer VFH, Eder P, Miehle K, et al. (2005) Interobserver variation for ultrasound determination of thyroid nodule volumes. *Thyroid* 15: 1169–1175. https://doi.org/10.1089/thy.2005.15.1169
48. Lee HJ, Yoon DY, Seo YL, et al. (2018) Intraobserver and interobserver variability in ultrasound measurements of thyroid nodules. *J Ultrasound Med* 37: 173–178. https://doi.org/10.1002/jum.14316
49. Özgen A, Erol C, Kaya A, et al. (1999) Interobserver and intraobserver variations in sonographic measurement of thyroid volume in children. *Eur J Endocrinol* 140: 328–331. https://doi.org/10.1530/eje.0.1400328

AIMS Press