Heliyon 10 (2024) e37924

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Video-based AI module with raw-scale and ROI-scale information for thyroid nodule diagnosis

Linghu Wu[a,1], Yuli Zhou[a,1], Mengmeng Liu[a], Sijing Huang[a], Youhuan Su[a], Xiaoshu Lai[a], Song Bai[a], Keen Yang[a], Yitao Jiang[c], Chen Cui[c], Siyuan Shi[c], Jinfeng Xu[a,**], Nan Xu[b,***], Fajin Dong[a,*]

[a] *Ultrasound Department, Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University, The First Affiliated Hospital, Southern University of Science and Technology), Shenzhen, 518020, Guangdong, China*

[b] *Division of Thyroid surgery, Department of General Surgery, Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University, The First Affiliated Hospital, Southern University of Science and Technology), Shenzhen, 518020, Guangdong, China*

[c] *Research and development department, Illuminate, LLC, Shenzhen, Guangdong, 518000, China*

ARTICLE INFO

Keywords:
Thyroid cancer
Artificial intelligence
Ultrasonography
Deep learning

ABSTRACT

Objectives: Ultrasound examination is a primary method for detecting thyroid lesions in clinical practice. Incorrect ultrasound diagnosis may lead to delayed treatment or unnecessary biopsy punctures. Therefore, our objective is to propose an artificial intelligence model to increase the precision of thyroid ultrasound diagnosis and reduce puncture rates.
Methods: We consecutively collected ultrasound recordings from 672 patients with 845 nodules across two Chinese hospitals. This dataset was divided into training, validation, and internal test sets in a ratio of 7:1:2. We constructed and tested six different model variants based on different video feature distillation strategies and whether additional information from ROI (Region of Interest) scales was used. The models' performances were evaluated using the internal test set and an additional external test set containing 126 nodules from a third hospital.
Results: The dual-stream model, which contains both raw-scale and ROI-scale streams with the time dimensional convolution layer, achieved the best performance on both internal and external test sets. On the internal test set, it achieved an AUROC (Area Under Receiver Operating Characteristic Curve) of 0.969 (95 % confidence interval, CI: 0.944–0.993) and an accuracy of 92.6 %, outperforming other variants (AUROC: 0.936–0.955, accuracy: 80.2%–88.3 %) and experienced radiologists (accuracy: 91.9 %). The AUROC of the best model in the external test was 0.931 (95 % CI: 0.890–0.972).
Conclusion: Integrating a dual-stream model with additional ROI scale information and the time dimensional convolution layer can improve performance in diagnosing thyroid ultrasound videos.

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: xujifeng@yahoo.com (J. Xu), xunan2005@163.com (N. Xu), dongfajin@szhospital.com (F. Dong).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2024.e37924
Received 12 May 2024; Received in revised form 11 September 2024; Accepted 13 September 2024
Available online 19 September 2024

1. Introduction

Thyroid cancer (TC) accounted for 586,000 cases and 44,000 deaths globally in 2020, ranking 9th in incidence and representing 3.8 % of all new cancer cases [1,2]. The detection rate of thyroid nodules (TNs) is increasing due to the widespread use of medical imaging technology, pathological biopsy, and medical monitoring [1,3]. Although only 7–15 % of all nodules are malignant [4], failing to diagnose and treat malignancies in time can lead to adverse outcomes such as abnormal thyroid function and metastasis. Therefore, early identification of pathological changes is crucial.

Currently, pathological biopsy is the recognized standard for diagnosing TNs [5]. However, its invasiveness is a significant barrier to making it a routine procedure. Pathological results can only be obtained through surgery or fine-needle aspiration (FNA) [6], which can cause complications such as hemorrhage, infection, injury to the recurrent laryngeal nerve or parathyroid gland, or even implantation and metastasis of tumor cells. The debate over the necessity of invasive procedures is ongoing [7]. Consequently, the adoption of conservative and risk-tailored management strategies has emerged as a significant issue [8–11].

Ultrasound has become the preferred method for examining the thyroid due to its non-invasive, convenient, low-cost, and real-time features. However, the potential for misdiagnosis arises from an overreliance on the radiologist's experience [12]. Objective approaches and increased accuracy are urgently needed.

Artificial intelligence has been considered as an effective method to improve the accuracy of thyroid ultrasound diagnosis [13]. Convolutional neural networks (CNNs) can automatically recognize and extract image features, which eliminates subjectivity in the diagnostic process [14]. They can be used in multiple medical tasks, including lesion detection, disease classification, and anatomical structure segmentation [15]. Previous studies have demonstrated the potential of artificial intelligence (AI) in assisting medical diagnosis. Some studies have shown that CNN-based methods for detecting TNs can achieve over 97.5 % accuracy [16,17] and 0.985 AUROC (area under the receiver operating characteristic curve) [18]. Experts have also demonstrated that CNN-based methods for detecting and classifying TNs are no less effective than experienced clinicians [19–21]. For example, Liu et al. [17] showed significantly improved sensitivity (0.964 vs. 0.928), specificity (0.780 vs. 0.366), and accuracy (0.928 vs. 0.816) using AI compared to clinicians.

Previous research on AI recognition of TNs primarily focused on image-level inference, generating predictions based on physician-selected images [22]. This approach suggests that the AI's acquired knowledge might be constrained by the subjectivity inherent in image selection, which could lead to the exclusion of vital information present in frames that are disregarded by the operators. Therefore, our aim was to develop an AI model capable of autonomously processing ultrasound videos and predicting malignancy. Chen et al. proposed a framework for video classification using a CNN as the backbone and a max-pooling layer to reduce information in the time dimension [23]. This demonstrated that pooling strategies could effectively link the knowledge from single image content to the entire recording in ultrasound-related tasks. In our study, we further explored information distillation strategies in the time dimension, proposing a method that employs a time-dimensional convolution layer to adjust the weight of each frame and generate final video-level predictions. Meanwhile, merging additional information from different modalities or scales may improve diagnostic performance. For example, utilizing the relationship between local and global regions can improve accuracy in pose estimation [24–26]. When it comes to video action classification, a two-stream architecture has been considered a classic implementation, using both raw stream and optical flow stream [27]. In our research, we considered both the raw-scale images and the ROI-scale fragments to be essential elements of ultrasound recordings. Consequently, we integrated the ROI-scale stream as an auxiliary stream to develop a dual-stream video classification neural network, which we then validated using both internal and external test datasets.

2. Materials and methods

2.1. Study population recruitment workflow

This was a prospective study that used US video sets. The inclusion criteria were: 1) patients aged over 18 years old; 2) with TNs ranging from 2 mm to 88 mm; 3) without any preoperative operation; 4) willingness to undergo thyroid surgery or FNA. The exclusion criteria were: 1) with any preoperative treatment, including FNA, thyroid surgery or histological investigation; 2) with poor-quality US videos. This study was approved by the local Research Ethics Committee of Shenzhen People's Hospital (Approval number: LL-KY2021-026-01). The written informed consent was obtained from all individual participants and all data used were acquired with institutional review board-approved protocols.

2.2. Process of US videos acquisition and analysis

The US examination was performed by two radiologists with expertise of 3–5 years (junior), two with 5–10 years (senior) and two experienced radiologists with more than 10 years experiments. US videos for TNs were captured by US systems with probes of 5–13 MHz (Mindray I9 and Sonoscape P60) and stored in DICOM format. Both longitudinal and transverse planes of the TNs were obtained. Diagnosis of benign and malignant TNs was provided by the three groups of radiologists independently. The ACR Thyroid Imaging Reporting and Data System (ACR TI-RADS) guideline [4], including the maximum diameter, composition, echogenicity, shape, margin, echogenic foci, and TI-RADS stages, were referred to evaluate the malignancy risk of each TN by two experienced radiologists. The ACR TI-RADS criteria for determining benign or malignant nodules relied on whether a patient underwent FNA biopsy or not. Discussions were conducted to obtain consensus results when inconsistencies existed.

2.3. Data organization and pre-processing

A total of 845 nodules, comprising 251 malignant and 594 benign cases from Shenzhen People's Hospital and Zhejiang Cancer Hospital, were utilized to train, validate, and test the classification modules by a distribution ratio of 7:1:2, respectively. To ascertain the robustness of our proposed model, an additional external test set was established, consisting of 126 nodules from 92 patients at the Longhua Branch of Shenzhen People's Hospital. Each video in our dataset contains imagery of a single nodule. The process of sample selection is illustrated in Fig. 1.

For each ultrasound screening included in this study, we cropped the main ultrasound window from the whole recorded video to eliminate information about patients and devices. Then, we converted the video to grayscale, utilized zero-padding, and resized each frame to a shape of 256 × 256. For each video, 64 frames were subsampled at equal intervals, and the value of each pixel was divided by 255 to obtain an input tensor with a fixed shape of 64 × 1 × 256 × 256, with values ranging from 0 to 1.

2.4. Algorithm design and model construction

The whole model consists of three main modules: the ROI detection module, the video feature extraction module, and the classification module. Raw videos are the inputs for the dual-stream model and are processed by the ROI detection module to crop out the ROI video. Subsequently, both raw video and ROI video are sent to the video feature extraction module, which processes each frame in parallel and generates feature maps. After temporal feature distillation, we obtain the final video-level features, merging both raw scale and ROI scale information. A simple fully connected layer is used as the final classification module to generate the video classification results (Fig. 2).

In clinical practice, physicians care about the specific sign of malignancy at the region near the lesion, such as calcification, margin, and composition. A common practice helping AI to pay attention is locating the nodule's region of interest (ROI) and letting the AI learn from ROIs. In order to realize it, we used a Faster-RCNN based object detection module to locate thyroid nodule of each frame. ROIs were cropped and resized to 128 × 128. An ROI video is constructed by integrating all the frames together. If nodule is not found in a frame, we will use an all-zero array to replace it.

The prior knowledge of distinguishing between benign and malignant lesions acquired from single-frame images is crucial for video processing. Carreira et al. produced an Inception-ResNet backbone based video classification model, which inherited the image level pretrained knowledge and achieved 80.2 % accuracy on HMDB-51 and 97.9 % on UCF-101 [28]. Chen et al. evaluated the performances of different backbones in thyroid ultrasound image classification task, and Inception-ResNetV2 achieved the highest AUROC of 0.94. In our study, we pretrained an inception-ResNetV2-based image-level backbone on a thyroid dataset containing 19,341 images of 7236 patients (2982 malignant patients) from Zhejiang Cancers Hospital. This backbone uses single grayscale image as the input and predicts its malignancy.

When processing video format data, we use the pretrained backbone to parallelly extract the features of each frame. Each feature map has the same shape of 1 × 1 × 1536 (1536 is the feature dimension defined by the image-level module of Inception-ResNetV2). In total, 64 feature maps will be created and used for downstream analysis. The last key component in video feature extraction module is the temporal distillation layer. Since all frames are used to produce features, the information contained in the final feature maps is huge and redundancy. Therefore, distillate important features to reduce the complexity of video level features is needed. Chen et al. used MaxPooling layer to distillate information, which means that only the most significant benign or malignant signal in each feature dimension (1536 dimensions in total) are used to make a prediction [23]. To improve the final accuracy, we also proposed two additional methods for temporal distillation, including AvgPooling layer and temporal convolution layer. In AvgPooling layer, we

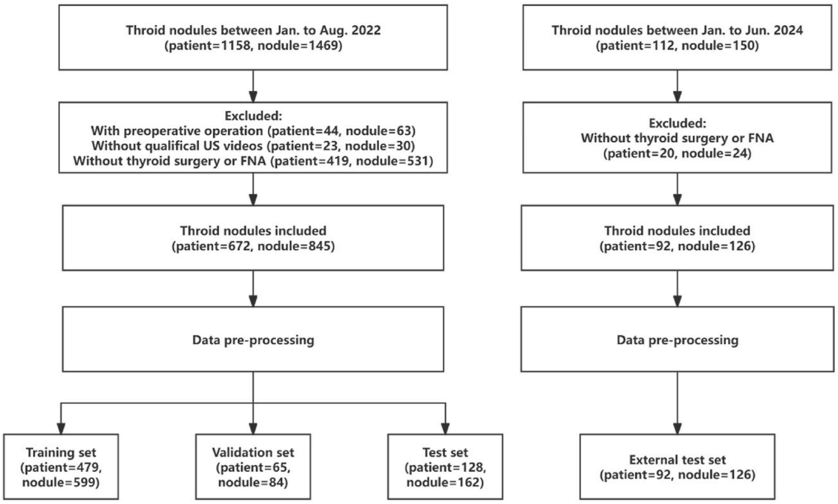
Throid nodules between Jan. to Aug. 2022
(patient=1158, nodule=1469)
Excluded:
With preoperative operation (patient=44, nodule=63)
Without qualifical US videos (patient=23, nodule=30)
Without thyroid surgery or FNA (patient=419, nodule=531)
Throid nodules included
(patient=672, nodule=845)
Data pre-processing
Training set
(patient=479, nodule=599)
Validation set
(patient=65, nodule=84)
Test set
(patient=128, nodule=162)
Throid nodules between Jan. to Jun. 2024
(patient=112, nodule=150)
Excluded:
Without thyroid surgery or FNA
(patient=20, nodule=24)
Throid nodules included
(patient=92, nodule=126)
Data pre-processing
External test set
(patient=92, nodule=126)

Fig. 1. Flowchart of inclusion and exclusion of the study subjects.

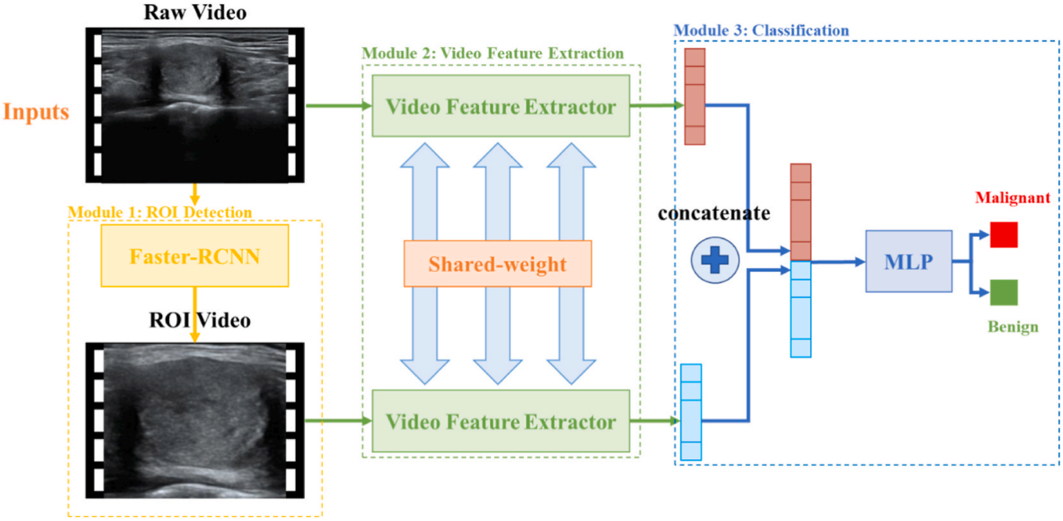

Fig. 2. Architecture of dual-stream neural network. The whole model consists of the ROI detection module, video feature extraction module, and classification module.

considered each frame share the same importance in prediction and the final video feature is the average result of all frames' features. In temporal convolution layer, we believed that the importance of different frames varies along the timeline. Therefore, we used a convolution kernel with the size of (64,1) to process the merged feature map (size: 64 × 1536) to generate the final video feature (size: 1 × 1536). The parameters in temporal convolution layer were learnt during the training process (Fig. 3).

After getting the final video feature map, we concatenated features from different scales and sent it to a fully connect layer. With the final fully connect layer, the model will provide a final two-classes prediction. We used data augmentation to mimic the image transformation in clinical data acquisition, including random stiff transformations, zoom, flip and rotation. Models were constructed using Pytorch 2.1.0 and Python 3.11 and trained on a server containing two Nvidia RTX 3090 GPUs. Batch size was set to 1 for 50 epochs with a learning rate of 0.001. Early stopping strategy was used and set to stop training when validation loss of five epochs does not decrease.

2.5. Evaluation indicators

To measure the performance of the trained network, common evaluation metrics were used. The primary endpoint of the study was the AUROC for TNs diagnosis, while the secondary endpoints were accuracy, sensitivity, and specificity. Ultrasonic videos were used to compare the performance of six different model variants as well as the performance of three radiologists with varied experiences.

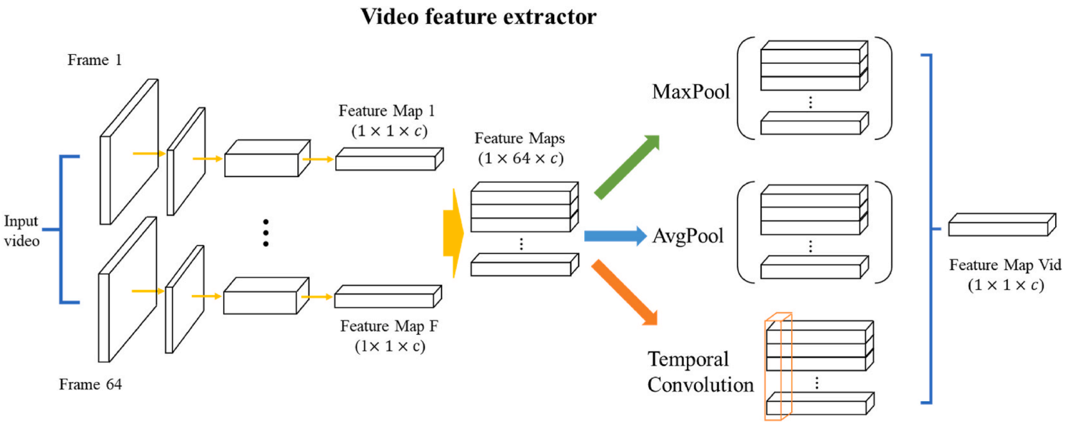

Fig. 3. Architecture of Video feature extractor. The video feature extractor utilizes knowledge from ultrasound frames, breaking down videos into 64-frame intervals. Each frame undergoes convolutional neural network processing, resulting in individual feature maps (1 × 1 × c) inherited from the image-level module. Different feature distillation strategies can be used to generate final video feature map, including MaxPool, AvgPool and Temporal Convolution.

3. Results

3.1. General and ultrasonic characteristics analysis

A total of 1158 participants were enrolled from January to August 2022. All participants underwent US examination, and US videos were collected. 44 patients were excluded because of prior thyroid procedures, while 23 patients were excluded due to poor-quality US videos. Additionally, 419 patients were excluded due to the lack of FNA biopsy or thyroid surgery. The study consisted of 672 patients (46.63 ± 13.32 years of age, range between 19 and 87 years old, 173 males (25.8 %) and 499 females (74.2 %) with 845 nodules in the final analysis. The patients' characteristics are listed in Table 1. Of the 845 nodules, 594 (70.30 %) were benign and 251 (29.70 %) were malignant. A comprehensive overview of the characteristics of these nodules is shown in Table 2. To further ascertain the robustness of our proposed model, we incorporated an external test set, which included 126 nodules from 92 patients. This set was collected between January and August 2022 at the Longhua Branch of Shenzhen People's Hospital.

3.2. Diagnostic performance of different variants on internal test set

The performance of six AI models for differentiating TNs was summarized in Table 3. The dual-stream model with temporal convolution layer achieved the highest AUROC of 0.969 (95 % CI: 0.944–0.993) with the highest accuracy of 0.926, outperforming all other models. In terms of using different feature distillation strategies, model with temporal convolution layers show better performance. When it comes to whether using additional ROI-scale information, dual stream variants always show higher AUROC.

3.3. Model performance compared with radiologists

The diagnostic performance in differentiation from malignant to benign nodules of test set by radiologists with different levels was showed in Fig. 4 and Table 3. When independently evaluating the TNs, the diagnosis of experienced radiologists showed higher accuracy, specificity, AUROC than the junior. The sensitivity of US diagnosis by experienced radiologists was also better than that by junior radiologists. The accuracy of the dual-stream temporal convolution model in TNs diagnosis is 0.926, which is significantly higher than that of junior and senior radiologists. The best model also shows an equivalent performance to experienced radiologists, with slightly higher sensitivity (0.900 vs. 0.884) and a similar specificity (0.938 vs. 0.935).

We further analyzed the mechanisms behind this phenomenon using Grad-CAM visualization. In Fig. 5, Case 1 depicts an image of papillary carcinoma, which both the dual-stream and single-stream models accurately diagnose, demonstrating a very similar emphasis on image features. However, when confronted with more complex malignant cases, such as the thyroid micro-papillary carcinomas presented in Case 2 and Case 3, the single-stream network's focus areas tend to stray from the actual nodular regions, thereby diminishing its diagnostic accuracy.

3.4. Ablation experiment

To evaluate whether using video format as raw input benefited AI diagnosis, we designed an ablation experiment that using pure image level Inception ResNetV2 model to make predictions on the images selected from internal test set. In this experiment, two radiologists were asked to pick the frame with most significant malignancy signatures in each video for image level prediction. On internal test set, the image level model reached an AUROC of 0.885 (95 % CI: 0.836–0.935). This result indicates that any video-based model has the potential to outperform models that operate solely at the single image level.

3.5. External test results

To test the robustness of our proposed model, we collected additional 126 thyroid ultrasound videos from Longhua Branch of Shenzhen People's Hospital. We tested six AI variants in external test set and dual stream model with temporal convolutional layer still achieved the highest AUROC of 0.931 (95 % CI: 0.890–0.972), with an accuracy over 0.85 (Table 4).

Table 1
The characteristics of patients.

	Total patients	Female	Male
No.	672	499	173
Age	46.63 ± 13.32	45.35 ± 12.65	50.38 ± 14.51
Multifocal/Unifocal	102/570	76/423	26/147
With/without HT	135/537	86/413	49/124

Multifocal: Number of patients with more than one nodule.
Unifocal: Number of patients with only one nodule.
HT: Hashimoto's thyroiditis.

Table 2
Demographic data of 845 nodules.

	Total nodules	Benign nodules	Malignant nodules
No. of nodules	845	594	251
Nodule size (mm)	13.36 ± 11.56	14.08 ± 12.67	11.65 ± 8.17
ACR TI-RADS			
1	140	140	0
2	180	175	5
3	110	109	1
4	184	145	39
5	231	25	206
Pathological type			
Nodular goiter	561	561	/
Follicular adenoma	20	20	/
Thyroid adenoma	8	8	/
Hashimoto's thyroiditis (non-nodule)	5	5	/
PTC	246	/	246
FTC	2	/	2
MTC	3	/	3

Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis
Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis
Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis
Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis
Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis
Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis
Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis
Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis
Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis
Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis
Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis
Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis	Video-based AI Model with Raw-Scale and ROI-Scale Information for Thyroid Nodule Diagnosis

Normally distributed numerical variables are shown by mean ± standard deviation. PTC: Papillary thyroid carcinoma; FTC: Follicular thyroid carcinoma; MTC: Medullary thyroid carcinoma.

Table 3
The diagnostic performances of the AI variants and radiologists on internal test set.

Variants	AUROC	95 % CI	Accuracy	Sensitivity	Specificity	PPV	NPV
Dual-Stream, with Temporal Convolution Layer	0.969	0.944–0.993	0.926	0.900	0.938	0.865	0.955
Single-Stream, with Temporal convolution layer	0.946	0.914–0.977	0.802	0.960	0.732	0.615	0.976
Dual-Stream, with MaxPool Layer	0.948	0.917–0.979	0.833	0.860	0.821	0.683	0.929
Single-Stream, with MaxPool Layer	0.936	0.899–0.972	0.864	0.900	0.848	0.726	0.950
Dual-Stream, with AvgPool Layer	0.955	0.924–0.985	0.858	0.940	0.821	0.701	0.968
Single-Stream, with AvgPool Layer	0.946	0.913–0.979	0.883	0.860	0.893	0.782	0.935
radiologists							
experienced		\	0.919	0.884	0.935	0.863	0.945
senior	\	\	0.867	0.837	0.880	0.766	0.920
Junior	\	\	0.830	0.697	0.891	0.750	0.863

4. Discussion

The prevalence of TC is increasing, yet the overdiagnosis and overtreatment of TC have become a significant concern. This concern arises from the high puncture rates, but low malignancy rates, leading to debates over subsequent treatment strategies for TC. Even non-operative invasive procedures like FNA are now being approached with increased caution. Consequently, there is an immediate need for an effective and objective method to reduce the risks associated with invasive examinations. While ultrasound examination

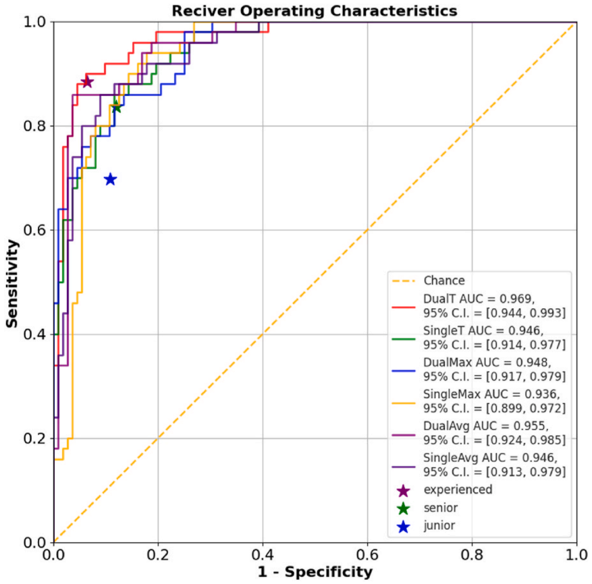

Fig. 4. Model performance compared with human radiologists. Multiple receiver operating characteristic curve illustrate the performance of various models in TNs diagnosis. The dual-stream temporal convolution model achieved an AUC of 0.969 in diagnosing thyroid nodules, significantly outperforming other models and human radiologists. The purple (★) labeled with "experienced" represents the diagnostic results of two experienced radiologists, the green (★) labeled with "senior" represents the diagnostic results of two senior radiologists, and the blue (★) labeled with "junior" represents the diagnostic results of two junior radiologists.

has become the preferred non-invasive diagnostic method, its diagnostic accuracy is hindered by inter-observer variability, necessitating improvement.

In this study, we have developed artificial intelligence models designed for the automatic and precise diagnosis of TNs using non-invasive ultrasonic videos. Through comparative analysis of various model variants, we have validated our hypothesis that incorporating additional information from ROI-scale streams can enhance diagnostic performance. This finding aligns with observations in other AI applications, where the integration of global and local information at different scales has been shown to boost model efficacy [29–31]. The ROI stream in the dual-stream network maintains a close focus on the nodule, while the raw stream takes into account both the nodule and the surrounding tissue. A notable drawback of the single-stream model is its reduced specificity, as evidenced by a score of 0.732 in our internal test set.

Analysis of research data demonstrates that the additional ROI scale information benefits final performance. Our analysis of the research data confirms that the inclusion of supplementary ROI-scale information significantly enhances overall performance. With the AI models we proposed, we achieved diagnostic accuracy, sensitivity, and specificity in TN diagnosis that are comparable to, if not superior, those of senior radiologists. Moreover, the dual-stream temporal convolution model exhibited a level of diagnostic accuracy, sensitivity, and specificity in TN diagnosis that matches that of experienced radiologists. Although our study presents promising outcomes for the automation of thyroid ultrasound diagnosis using AI, there are areas that warrant further investigation. For instance, the video data in this study were obtained by well-trained physicians adhering to a standard operating procedure. In other medical facilities with different protocols, their ultrasound examinations might present different characteristics, particularly in terms of the temporal dynamics of the image sequences. Additionally, our research has primarily concentrated on papillary thyroid carcinoma, which is the most prevalent pathological subtype. Therefore, the generalizability of our findings to all subtypes of TC is yet to be determined. In future research, we intend to expand our dataset to include more pathological types of TC, with the goal of refining our model's ability to accurately diagnose each TC subtype and thereby extending its clinical utility.

In conclusion, our study introduces a dual-stream model equipped with temporal convolution for the automated and precise diagnosis of thyroid nodules utilizing non-invasive ultrasound videos. The findings of our research substantiate the notion that leveraging information across various scales can significantly improve the predictive capabilities of AI in medical diagnostics.

Financial support

No.

Guarantor

The scientific guarantor of this publication is Fajin Dong.

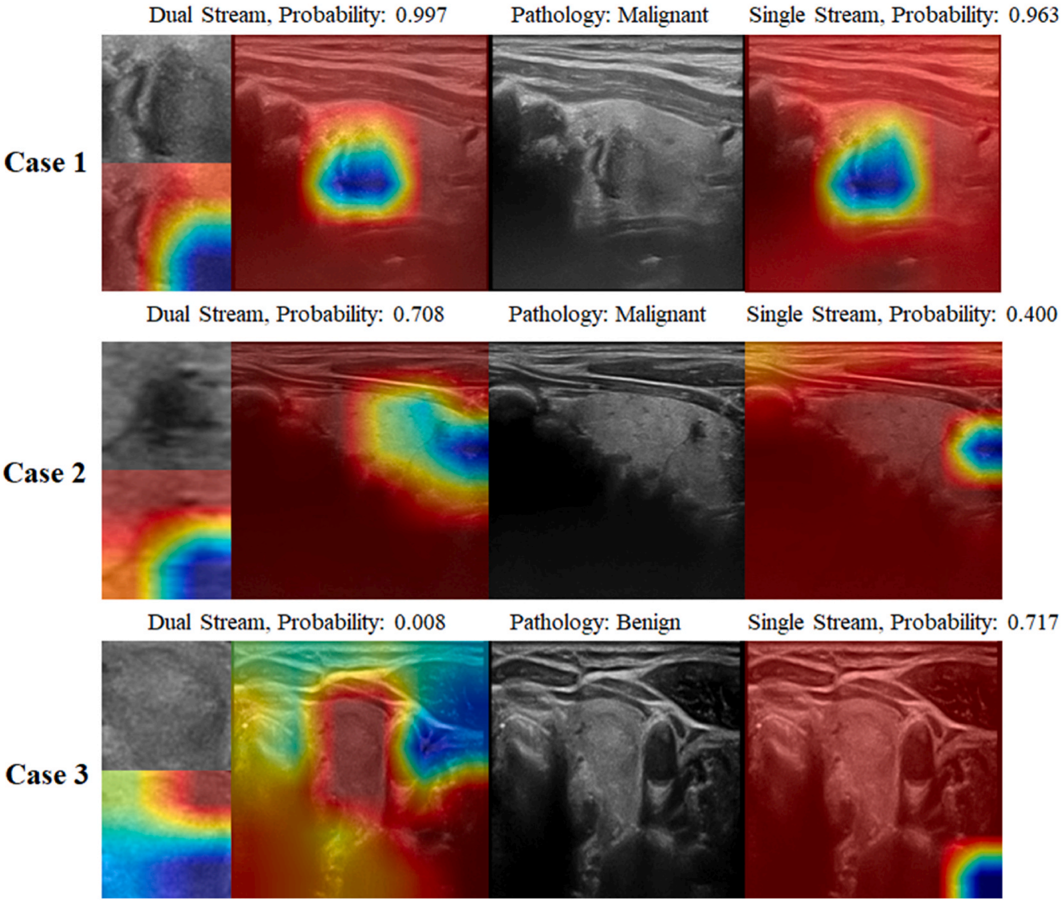
Dual Stream, Probability: 0.997
Pathology: Malignant
Single Stream, Probability: 0.963
Case 1
Dual Stream, Probability: 0.708
Pathology: Malignant
Single Stream, Probability: 0.400
Case 2
Dual Stream, Probability: 0.008
Pathology: Benign
Single Stream, Probability: 0.717
Case 3

Fig. 5. Heatmap analysis using Grad-CAM. For easily distinguishable malignant nodules, both single-stream and dual-stream networks focus on the same features and predict a high probability of malignancy (Case 1). For relatively hard-to-distinguish malignant nodules, the ROI stream of the dual-stream network focuses on the nodule area and predicts a malignancy probability of 0.708, whereas the raw scale stream, which attends to surrounding relevant regions, predicts a lower malignancy probability of 0.400 (Case 2). In benign nodules, the dual-stream network effectively captures both the nodule and its surrounding features, predicting a malignancy probability of 0.008. In contrast, the single-stream network may focus on irrelevant areas, resulting in a higher malignancy probability of 0.717 and leading to potential misjudgments (Case 3).

Table 4
The diagnostic performances of the AI variants on external test set.

Variants	AUROC	95 % CI	Accuracy	Sensitivity	Specificity	PPV	NPV
Dual-Stream, with Temporal Convolution Layer	0.931	0.890–0.972	0.857	0.935	0.781	0.806	0.926
Single-Stream, with Temporal convolution layer	0.912	0.864–0.960	0.817	0.839	0.797	0.800	0.836
Dual-Stream, with MaxPool Layer	0.913	0.866–0.960	0.833	0.758	0.906	0.887	0.795
Single-Stream, with MaxPool Layer	0.893	0.840–0.946	0.810	0.806	0.813	0.806	0.813
Dual-Stream, with AvgPool Layer	0.921	0.873–0.968	0.865	0.952	0.781	0.808	0.943
Single-Stream, with AvgPool Layer	0.870	0.811–0.929	0.762	0.790	0.734	0.742	0.783

Statistics and biometry

No complex statistical methods were necessary for this paper.

Informed consent

Written informed consent was obtained from all subjects (patients) in this study.

Ethical approval

This study was approved by Shenzhen People's Hospital (approval number: LL-KY2021-026-01).

Code and data availability

To facilitate the reproduction of our work and support further research in thyroid ultrasound, we have published our code on GitHub, along with the internal test set. Anybody can access the repository using this link: https://github.com/dteamsz/ThyroidDualStream.git.

CRediT authorship contribution statement

Linghu Wu: Writing – original draft, Data curation, Conceptualization. **Yuli Zhou:** Project administration. **Mengmeng Liu:** Formal analysis. **Sijing Huang:** Supervision. **Youhuan Su:** Data curation. **Xiaoshu Lai:** Investigation. **Song Bai:** Resources, Data curation. **Keen Yang:** Methodology. **Yitao Jiang:** Validation, Methodology. **Chen Cui:** Formal analysis, Data curation. **Siyuan Shi:** Methodology, Conceptualization. **Jinfeng Xu:** Writing – review & editing, Supervision, Data curation. **Nan Xu:** Project administration, Data curation. **Fajin Dong:** Writing – review & editing, Resources, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This project was supported by Shenzhen Science and Technology Program(GJHZ20220913142801003) and Clinical Research Project of Shenzhen People's Hospital (SYLCYJ202118).

References

[1] H. Sung, J. Ferlay, R.L. Siegel, et al., Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, Ca - Cancer J. Clin. 71 (3) (2021) 209–249, https://doi.org/10.3322/caac.21660.

[2] W. Cao, H.D. Chen, Y.W. Yu, N. Li, W.Q. Chen, Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020, Chin. Med. J. (Engl). 134 (7) (2021) 783–791, https://doi.org/10.1097/CM9.0000000000001474.

[3] S. Guth, U. Theune, J. Aberle, A. Galach, C.M. Bamberger, Very high prevalence of thyroid nodules detected by high frequency (13 mhz) ultrasound examination, Eur. J. Clin. Invest. 39 (8) (2009) 699–706, https://doi.org/10.1111/j.1365-2362.2009.02162.x.

[4] G.L. Francis, S.G. Waguespack, A.J. Bauer, et al., Management guidelines for children with thyroid nodules and differentiated thyroid cancer, Thyroid 25 (7) (2015) 716–759, https://doi.org/10.1089/thy.2014.0460.

[5] S. Filetti, C. Durante, D. Hartl, et al., Thyroid cancer: esmo clinical practice guidelines for diagnosis, treatment and follow-updagger, Ann. Oncol. 30 (12) (2019) 1856–1883, https://doi.org/10.1093/annonc/mdz400.

[6] S. Vaccarella, S. Franceschi, F. Bray, C.P. Wild, M. Plummer, L. Dal Maso, Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis, N. Engl. J. Med. 375 (7) (2016) 614–617, https://doi.org/10.1056/NEJMp1604412.

[7] B.R. Haugen, E.K. Alexander, K.C. Bible, et al., 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) (2015) 1–133, https://doi.org/10.1089/thy.2015.0020, 2016.

[8] F. Pacini, M. Schlumberger, H. Dralle, R. Elisei, J.W. Smit, W. Wiersinga, European consensus for the management of patients with differentiated thyroid carcinoma of the follicular epithelium, Eur. J. Endocrinol. 154 (6) (2006) 787–803, https://doi.org/10.1530/eje.1.02158.

[9] F. Pacini, F. Basolo, R. Bellantone, et al., Italian consensus on diagnosis and treatment of differentiated thyroid cancer: joint statements of six Italian societies, J. Endocrinol. Invest. 41 (7) (2018) 849–876, https://doi.org/10.1007/s40618-018-0884-2.

[10] A.L. Mitchell, A. Gandhi, D. Scott-Coombes, P. Perros, Management of thyroid cancer: United Kingdom national multidisciplinary guidelines, J. Laryngol. Otol. 130 (S2) (2016) S150–S160, https://doi.org/10.1017/S0022215116000578.

[11] J.Y. Kwak, K.H. Han, J.H. Yoon, et al., Thyroid imaging reporting and data system for us features of nodules: a step in establishing better stratification of cancer risk, Radiology 260 (3) (2011) 892–899, https://doi.org/10.1148/radiol.11110206.

[12] S.H. Choi, E.K. Kim, J.Y. Kwak, M.J. Kim, E.J. Son, Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules, Thyroid 20 (2) (2010) 167–172, https://doi.org/10.1089/thy.2008.0354.

[13] Z. Akkus, J. Cai, A. Boonrod, et al., A survey of deep-learning applications in ultrasound: artificial intelligence-powered ultrasound for improving clinical workflow, J. Am. Coll. Radiol. 16 (9 Pt B) (2019) 1318–1328, https://doi.org/10.1016/j.jacr.2019.06.004.

[14] L. Wang, S. Yang, S. Yang, et al., Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the yolov2 neural network, World J. Surg. Oncol. 17 (1) (2019) 12, https://doi.org/10.1186/s12957-019-1558-z.

[15] Y.T. Shen, L. Chen, W.W. Yue, H.X. Xu, Artificial intelligence in ultrasound, Eur. J. Radiol. 1392021) 109717, https://doi.org/10.1016/j.ejrad.2021.109717.

[16] W. Song, S. Li, J. Liu, et al., Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition, IEEE J. Biomed. Health Inform 23 (3) (2019) 1215–1224, https://doi.org/10.1109/JBHI.2018.2852718.

[17] T. Liu, Q. Guo, C. Lian, et al. Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks, Med. Image Anal. 582019) 101555, https://doi.org/10.1016/j.media.2019.101555.

[18] J. Ma, F. Wu, T. Jiang, J. Zhu, D. Kong, Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images, Med. Phys. 44 (5) (2017) 1678–1691, https://doi.org/10.1002/mp.12134.

[19] Y.J. Kim, Y. Choi, S.J. Hur, et al., Deep convolutional neural network for classification of thyroid nodules on ultrasound: comparison of the diagnostic performance with that of radiologists, Eur. J. Radiol. 1522022) 110335 https://doi.org/10.1016/j.ejrad.2022.110335.

[20] X. Li, S. Zhang, Q. Zhang, et al., Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study, Lancet Oncol. 20 (2) (2019) 193–201, https://doi.org/10.1016/S1470-2045(18)30762-9.

[21] J. Ma, F. Wu, J. Zhu, D. Xu, D. Kong, A pre-trained convolutional neural network based method for thyroid nodule diagnosis, Ultrasonics 732017) 221-230 https://doi.org/10.1016/j.ultras.2016.09.011.

[22] C. Chen, Y. Jiang, J. Yao, et al., Deep learning to assist composition classification and thyroid solid nodule diagnosis: a multicenter diagnostic study, Eur. Radiol. 34 (4) (2024) 2323–2333, https://doi.org/10.1007/s00330-023-10269-z.

[23] J. Chen, Y. Jiang, K. Yang, et al., Feasibility of using ai to auto-catch responsible frames in ultrasound screening for breast cancer diagnosis, iScience 26 (1) (2023) 105692, https://doi.org/10.1016/j.isci.2022.105692.

[24] L. H, L. T, C. Y, Z. Z, F.L. Y, Ehpe: skeleton cues-based Gaussian coordinate encoding for efficient human pose estimation, IEEE Trans. Multimed. (2022) 1–12, https://doi.org/10.1109/TMM.2022.3197364.

[25] Z. C, L. H, D. Y, X. B, L. Y, Tokenhpe: learning orientation tokens for efficient head pose estimation via transformers, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8897–8906, https://doi.org/10.1109/CVPR52729.2023.00859.

[26] L. T, L. H, Y. B, Z. Z, Ldcnet: limb direction cues-aware network for flexible hpe in industrial behavioral biometrics systems, IEEE Trans. Ind. Inf. 20 (6) (2024) 8068–8078, https://doi.org/10.1109/TII.2023.3266366.

[27] K Simonyan, A Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Process. Syst. 27 (2014).

[28] C. J, Z. A, Quo vadis, action recognition? A new model and the kinetics dataset, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4724–4733, https://doi.org/10.1109/CVPR.2017.502.

[29] H. Liu, C. Zhang, Y. Deng, T. Liu, Z. Zhang, Y. Li, Orientation cues-aware facial relationship representation for head pose estimation via transformer, IEEE Trans. Image Process. 322023) 6289-6302 https://doi.org/10.1109/TIP.2023.3331309.

[30] L. H, L. T, Z. Z, K.S. A, Y. B, L. Y, Arhpe: asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction, IEEE Trans. Ind. Inf. 18 (10) (2022) 7107–7117, https://doi.org/10.1109/TII.2022.3143605.

[31] H. Liu, S. Fang, Z. Zhang, D. Li, J. Wang, Mfdnet: collaborative poses perception and matrix Fisher distribution for head pose estimation, IEEE Trans. Multimed. PP (99) (2021) 1.