

AI-enhanced ultrasound video analysis for BI-RADS 4-5 breast lesions: A multicenter study on diagnostic accuracy and biopsy reduction

Wei-Bin Li

Xiamen Key Laboratory of Endocrine-Related Cancer Precision Medicine, Xiamen Research Center of Clinical Medicine in Breast & Thyroid Cancers

Zhi-cheng Du

Xiamen Key Laboratory of Endocrine-Related Cancer Precision Medicine, Xiamen Research Center of Clinical Medicine in Breast & Thyroid Cancers

Wen-He Huang

Xiamen Key Laboratory of Endocrine-Related Cancer Precision Medicine, Xiamen Research Center of Clinical Medicine in Breast & Thyroid Cancers

Ri-Yu Qiu

School of Informatics, Xiamen University,

Yun-Chao Chen

Department of Ultrasound Medical Xiang'An Hospital of Xiamen University

Lai-Shun Lu

Department of Ultrasonic Medicine Shantou Central Hospital,

Ji-Yuan Zhong

Department of Ultrasonic Medicine Shantou Central Hospital,

Le-Hang Guo

Tongji University School of Medicine

Mei-Qing He

Department of Ultrasonic Medicine Shaan 'xi Provincial People's Hospital

Lian-Sheng Wang

School of Informatics, Xiamen University

Guo-Jun Zhang

zhangguojun@kmmu.edu.cn

Yunnan Cancer Hospital & The Third Affiliated Hospital of Kunming Medical University & Peking University Cancer Hospital
Yunnan

Research Article

Keywords: Ultrasound video image, Breast lesion, Computer-aided diagnosis, Benign and malignant, BI-RADS 4–5

Posted Date: May 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9498512/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Purpose: This study aims to develop a computer-aided diagnosis (CAD) model for aiding ultrasonographers in classification of benign or malignant breast lesions in BI-RADS categories 4-5 using ultrasound videos.

Materials and Methods: This prospective multicenter study enrolled 532 breast lesions categorized as breast imaging reporting and data system (BI-RADS) 4-5 by ultrasound. Of them, 334 breast lesions were included as training and validation sets. The remaining 198 lesions were assigned to a test set. We acquired both static and dynamic video ultrasound images of lesions for each patient and utilized them to develop a deep learning model. Temporal-guided multi-scale multi-instance learning network (TGMIL) with the video images was also developed. In this study, four metrics were calculated: area under the curve (AUC), accuracy, precision and sensitivity. In addition, pairwise DeLong and McNemar were used as statistical tests.

Results: A total of 120 patients with 198 videos (benign, 45.8 ± 11.5 years; malignant, 50.7 ± 11.0 years)

were evaluated. On the video-based test set, TGMIL achieved an AUC of 0.824 (95% CI, 0.790–0.857), significantly outperforming junior ultrasonographers (AUC, 0.698; 95% CI, 0.615–0.781; $p < 0.05$) and attending ultrasonographers (AUC, 0.808; 95% CI, 0.735–0.880; $p < 0.05$). With TGMIL assistance, diagnostic accuracy improved from 69.2% to 85.9% for juniors ($p < 0.05$) and from 81.8% to 87.3% for attendings ($p < 0.05$); sensitivity increased from 68.4% to 82.7% for juniors ($p < 0.05$) and from 80.8% to 86.5% for attendings ($p < 0.05$).

Conclusions: The TGMIL model significantly improved the diagnostic performance of junior and attending ultrasonographers for BI-RADS 4-5 breast lesions.

1. Introduction

Diagnostic accuracy of breast lesions preoperatively is challenging, and significantly impacts clinical decisions, i.e., to operate or not to operate, made by breast surgeons. To accurately distinguish benign from malignant lesions, clinicians and radiologists have made extensive efforts to improve lesion classification^[1]. For example, the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS®, 5th edition, American College of Radiology) provides standardized guidance for reporting and assessment^[2, 3].

Asian women have denser breast tissue compared to other ethnic groups^[4–6]. Particularly in China, the main and auxiliary imaging methods for breast cancer screening are mammography and ultrasound, as suggested by Guidelines for diagnosis and treatment of advanced breast cancer in China (2024 edition) *et al.*^[7]. Although mammography has been a mainstay in breast cancer screening, optimal accuracy in women with high-dense breast is difficult to achieve with screening by mammography alone^[8]. Breast ultrasound is a convenient diagnostic modality showing an important role in screening of individuals with dense breast tissue^[9]. Despite advances in ultrasound technology that have improved diagnostic performance^[10, 11], ultrasound screening still has limitations, including heavy dependence on operator experience and training^[12]. Suboptimal performance may result in missed small lesions or unnecessary core-needle biopsies and surgeries^[13].

Deep learning (DL) models^[14] show strong self-learning abilities and have been developed to assist ultrasonographers in diagnosing breast lesions^[15–21]. Machine learning (ML) methods could classify non-cystic benign and malignant breast lesions with higher accuracy^[22]. For instance, Pfoh *et al.* reported that ML algorithms combined with shear-wave elastography (SWE) led to a 50.3% reduction in unnecessary biopsies^[23]. Moreover, a convolutional neural network (CNN) trained to classify benign tumors, malignant tumors, inflammatory masses, and adenosis using US images achieved AUCs from 0.89 to 0.91 and an accuracy of 89.2% for the test set^[24].

However, in traditional image classification models, selecting a single frame or several frames from a video clip is a critical step during dynamic scanning^[25–26]. To date, studies on computer-aided diagnosis (CAD) analysis have focused primarily on

static (single frame) ultrasound images^[27–29], The ideal AI-based breast diagnosis model should outperform even highly experienced ultrasonographers in diagnosing suspicious lesions^[30]..

In this study, we develop a DL model based on a temporal-guided, multi-scale, multi-instance learning network (TGMIL) to achieve accurate diagnosis of BI-RADS 4–5 breast lesions.

2. Materials and methods

2.1. Study population and data sets

This prospective multicenter diagnostic study was approved by the Ethics Committee of Hospital A (XAHLL2022060) and informed consent was obtained from all participants. Breast ultrasound video data was acquired from four hospitals via transmission of images by cloud services or Baidu Netdisk (<https://meet.wonderview.cn:7809/#/login>; <https://yun.baidu.com/>) from November 2019 to January 2023.

The inclusion criteria for the study were as follows: (a) lesions on video images classified as BI-RADS category 4–5; (b) lesions clearly visible in grayscale images without measurement labels and in the color Doppler sampling window in video images; (c) lesions detected with a high-frequency probe and larger than 5 mm; (d) diagnoses confirmed by pathology after surgery or ultrasound-guided biopsy. The exclusion criteria were: (a) presence of a prosthesis on the affected side; (b) poor quality of traditional image classification models or videos; (c) missed pathological diagnosis; (d) unclear lesion type.

We performed a prospective multicenter diagnostic study using breast cancer US video data at Hospital A and Hospital B between November 2019 and July 2022. Finally, 334 breast lesions (61 benign lesions and 273 malignant lesions) were included as training and validation sets. An additional 198 lesions (50 benign and 148 malignant) at Hospital A, Hospital C, Hospital B and Hospital D collected between November 2020 and January 2023 were assigned to the test set, as shown in Fig. 1.

2.2. Ultrasound and pathology examinations

All breast ultrasound images were acquired using machines from GE Healthcare (Chicago, IL, USA), Siemens (Munich, Germany) or Philips (Amsterdam, the Netherlands). Image quality control was performed for both the training and test sets. The L12-5 ultrasound probe, a broadband linear probe with a frequency of 12 MHz, was used for dynamic scanning of lesions, from the first appearance to complete disappearance, with continuous scanning throughout the entire process, including traditional image classification models (such as color Doppler images). The scanning and ultrasound diagnosis of lesions was conducted by experienced ultrasonographers using the cross-plane screening method. Imaging data for both longitudinal and transverse views of each lesion were stored in the ultrasound equipment and subsequently transferred to a hard disk for storage in a picture archiving and communications system (PACS). All images were saved as traditional image classification models and dynamic videos, extracted from the ultrasound devices and converted into DICOM format. The process for image capture and storage standards for the ultrasound data is detailed in the Supplemental Material (Supplementary).

Tissue samples for pathological diagnosis were obtained through ultrasound-guided biopsy or surgical resection. For patients with multiple lesions, only the lesions with the largest diameter and the highest BI-RADS category were analyzed to ensure statistical independence. If the biopsy did not provide sufficient information for diagnosis, surgical resection was recommended. Pathological diagnoses were performed by pathologists with more than 8 years of experience in breast pathology.

2.3. Development of the ultrasound computer-aided diagnosis model

The development of the deep learning algorithm was composed of three steps: pre-processing, model design, and training.

2.3.1 Pre-processing step

The original video data in the training data set were processed through the three following steps: (a) The video was converted into the lossless format NPY through SimpleITK and NumPy, and redundant black edges and all text information were removed at the same time; (b) Each video was processed using the K-means algorithm to extract 15 key frames; (c) The video frame size was adjusted to $512 \cdot 512$ pixels, and no other image enhancement was performed; (d) The video was standardized to limit the pixel value to between 0 and 1.

2.3.2 Model design step

Each frame is encoded by a VGG-13 backbone. We initialise the network with publicly available pretrained weights and fine-tune all layers on ultrasound cine loops, a strategy that accelerates convergence and improves generalisability on the limited dataset. The result is an ordered sequence of frame embeddings, one channel per frame. After frame-level feature extraction, temporal information is encoded using a linear layer and concatenated with the frame features along the feature dimension. Subsequently, an attention mechanism derived from a MIL framework is employed to guide a temporally-aware attention process. This mechanism adaptively assigns weights across frames, selectively emphasizing diagnostically relevant frames and generating a global video-level representation. By automatically suppressing irrelevant and redundant information across frames, this method not only enhances the accuracy of video-based diagnosis but also provides interpretable results by highlighting the most informative frames.

2.3.3 Training step

In this work, the cross-entropy loss was calculated on the basis of the predicted label and real expression of each video and was applied to the optimization. The network was implemented using Pytorch on the basis of SGD training with the learning rate set to $5e-4$, and the batch size set to 1. The entire architecture was trained on one NVIDIA RTX 3090 GPU.

2.4 Performance validation

This study adopted a train-validation scheme and evaluated performance across four external test sets. Classification was assessed using AUC, accuracy, precision, and sensitivity, which respectively reflect overall discriminative ability, overall correctness, reliability of positive predictions, and the model's ability to identify true positives. To comprehensively assess the diagnostic performance of the proposed TGMIL model, we conducted three levels of comparison: (1) against other video-based deep learning models, (2) against traditional static image-based models, and (3) against experienced ultrasonographers under both unaided and AI-assisted conditions.

2.4.1 Comparison with other video-based models

We benchmarked TGMIL against several widely used video classification architectures, including 3D-CNN, CSN, and $R(2 + 1)D-50$. All models were trained and evaluated under the same data conditions using four standard classification metrics: AUC, accuracy, specificity, and sensitivity. Performance curves were plotted, and statistical comparisons were conducted to evaluate superiority across the full operating range.

2.4.2 Comparison with static image-based models

TGMIL was further compared with BU-Mamba, ResNet50, VGG16 and ViT B/32, which are static ultrasound image classifier. The same test set was used, and four evaluation metrics (AUC, accuracy, specificity, and sensitivity) were calculated for all models. This comparison aimed to highlight the diagnostic benefit of leveraging temporal context in dynamic ultrasound data.

2.4.3 Diagnostic performance of benign and malignant breast lesions compared between CAD model and ultrasonographers on both traditional image classification models or videos

A test set consisting of 198 ultrasound videos was reviewed by nine ultrasonographers of varying levels of experience (juniors, less than 3 years of experience, $n = 3$; attendings, 4 to 9 years of experience, $n = 3$; and seniors, ≥ 10 years of experience, $n = 3$). Each ultrasonographer reviewed a different case order, and the pathology results were unknown at the time

of assessment. Diagnostic performance was examined in four steps. Step 1: Ultrasonographers assessed the 198 lesions on images without TGMIL assistance. Step 2: Four weeks later, the case sequence was changed again and the ultrasonographers assessed 198 lesions using ultrasound videos with TGMIL assistance. The ultrasound diagnostic criteria for benign and malignant breast lesions followed the 5th edition BI-RADS guideline, including shape, orientation, margin, posterior feature, and echo pattern^[31]. The BI-RADS classification results were recorded with benign or malignant probabilities^[32]. Subsequently, we performed the following analyses: we compared the diagnostic performance of TGMIL and the ultrasonographers. We compared the diagnostic performance of TGMIL and the ultrasonographers using a routine method for reviewing ultrasound traditional image classification models and videos, stratified by BI-RADS category subgroups.

2.5 Comparison between TGMIL and Traditional Image Classification Models

Presents a detailed comparison between the proposed TGMIL model and BU-Mamba^[33], a representative static image-based classification model, across four commonly used evaluation metrics: AUC, Accuracy, Specificity, and Sensitivity.

2.6 Statistical analysis

All statistical analyses were performed using SPSS software for Windows (v.20.0) and Python software (v.3.8.5150.0; <http://www.python.org>). Chi square tests and *t*-tests were used to compare the differences between patients with benign and malignant lesions. Normally distributed data were represented as mean $\pm$ standard deviation ($\bar{x} \pm s$). Data with non-normal distributions and unequal variance were represented by the median (interquartile range) and expressed as proportions. The McNemar test was used to compare the differences in sensitivity, specificity, and accuracy between TGMIL and the ultrasonographers. A cut-off value 0.5 was used for TGMIL classification. Differences were considered statistically significant at $P < 0.05$. Python software was employed for AI-based imaging analysis and diagnostic tests. GraphPad Prism software (v.8.0.2) was applied to generate receiver operating characteristic (ROC) curves and calculate the area under the curve (AUC). The heatmap on the right presents pairwise DeLong test *p*-values comparing AUCs between centers.

3. Results

3.1 The computer-aided diagnosis model: Temporal-guided multi-scale multi-instance learning network (TGMIL)

The TGMIL model's curve is consistently positioned above and to the left of the curves of the other comparison methods. This visually indicates that at the same False Positive Rate, the TGMIL model achieves a higher True Positive Rate. Specifically, the TGMIL model achieved an AUC value of (0.824, 95% CI: 0.790–0.857), which is significantly higher than that of the 3D-CNN model (AUC value of 0.665), the CSN model (AUC value of 0.613), and the R (2 + 1) D-50 model (AUC value of 0.678). This demonstrates a clear advantage of the TGMIL model in terms of overall discriminative ability (Fig. 2).

3.2 Diagnostic performance of the TGMIL and Traditional Image Classification Models

We compared the diagnostic performance of the proposed TGMIL model against four representative image-based classification models: BU-Mamba, ResNet50, VGG16, and ViT B/32, using four commonly adopted metrics: AUC, accuracy, precision, and sensitivity (Table 1). The TGMIL model consistently achieved superior performance across all metrics. Specifically, it obtained the highest AUC (0.824, 95% CI: 0.758–0.889) indicating strong discriminative capability in distinguishing benign from malignant lesions. In contrast, ViT B/32 yielded the lowest AUC (0.656, 95% CI: 0.476–0.836), suggesting limited ability to generalize across decision thresholds. In terms of accuracy, TGMIL also outperformed other models (0.833, 95% CI: 0.784–0.883), reflecting the highest overall classification correctness. Although BU-Mamba achieved high sensitivity (0.988, 95% CI: 0.984–0.993), its precision was relatively low (0.752, 95% CI: 0.748–0.757), implying a

tendency toward over-diagnosis and a higher false-positive rate. In comparison, TGMIL balanced both high sensitivity (0.966, 95% CI: 0.936–0.996) and high precision (0.836, 95% CI: 0.782–0.891), indicating reliable identification of malignant lesions while maintaining specificity. Other models such as ResNet50 and VGG16 exhibited relatively competitive sensitivity and precision, but their wider confidence intervals indicated greater variability and lower robustness. Overall, the results demonstrate that TGMIL provides more accurate, consistent, and clinically meaningful predictions than traditional static image-based approaches.

Table 1

Participant characteristics and ultrasound videos in the training, validation and test set. [†]*P*: Participant characteristics of training set between benign and malignant alone. [‡]*P*: Participant characteristics of test set between benign and malignant alone. [§]*P*: Comparison of training set and test set.

	Training and validation set			Test set			[†] <i>P</i>	[‡] <i>P</i>	[§] <i>P</i>
	Benign	Malignant	Σ	Benign	Malignant	Σ			
Patients (n, %)	42 (18.5)	185 (81.5)	227	34(28.3)	86 (71.7)	120			
Videos (n, %)	61 (18.3)	273 (81.7)	334	50(25.2)	148 (74.8)	198			
Breast nodules	61 (18.3)	273 (81.7)	334	50(25.2)	148 (74.8)	198			
Age(years)									
Mean	42.9 ± 10.6	52.1 ± 11.4		45.8 ± 11.5	50.7 ± 11.0		<0.05	<0.05	0.181
Range	22–68	26–87		19–68	26–77				
Nodule size (mm)									
Mean	18.2 ± 7.8	21.7 ± 8.6		17.1 ± 8.6	23.1 ± 8.6		0.05	<0.05	0.511
Range	6–48	8–48		5–47	7–49				
BI-RADS Category									
4a	45	34		39	21				
4b	15	141		10	63				
4c	1	60		1	49				
5	0	38		0	15				
Histologic type									
Malignant tumor									
Invasive breas carcinoma	NA	167		NA	89				
Carcinoma of mixed type	NA	56		NA	30				
Ductal carcinoma in situ	NA	38		NA	18				
Other malignant tumors	NA	12		NA	11				
Benign tumor									
Fibroadenoma	34	NA		28	NA				
Papilloma	10	NA		9	NA				
Other benign tumors	17	NA		13	NA				

Patients with malignant lesions were larger than benign lesions ($P < 0.05$), though no significant difference was observed in age or lesion size between the training and test sets ($P > 0.05$), as shown in Table 2.

Table 2
Comparison of AUC, Accuracy, Precision, and Sensitivity Between TGMIL and Static Image Models..

BU-Mamba	AUC	Accuracy	Precision	Sensitivity
	0.703 ± 0.011	0.766 ± 0.009	0.752 ± 0.002	0.988 ± 0.002
ResNet50	0.688 ± 0.089	0.800 ± 0.058	0.802 ± 0.061	0.967 ± 0.031
VGG16	0.718 ± 0.090	0.770 ± 0.058	0.821 ± 0.061	0.882 ± 0.052
ViT B/32	0.656 ± 0.092	0.757 ± 0.060	0.795 ± 0.064	0.903 ± 0.046
TGMIL	0.824 ± 0.034	0.833 ± 0.025	0.836 ± 0.028	0.966 ± 0.015

3.3 Diagnostic performance of the model and the four external test sets

As shown in Fig. 4, the model demonstrated excellent performance on the training set, with a 0.985, CI: 0.984–0.998), indicating strong fitting capacity on seen data. As shown in Fig. 3 and Table 3, we report the model's performance in terms of AUC, accuracy, specificity, and sensitivity on the validation set, and four external test centers to provide a comprehensive evaluation of its classification ability. On the validation set, performance slightly declined but still reflected good generalization, with a specificity of 0.877 and sensitivity of 0.906, suggesting the model remains effective at identifying positive cases. Across the four external test sets, the model showed stable and robust performance. Hospital B achieved the highest AUC (0.886, 95% CI: 0.815–0.958) and specificity was 0.911, although its sensitivity was relatively lower 0.769, suggesting a more conservative prediction behavior for positive cases. In contrast, both Hospital D and Hospital A reached perfect sensitivity 1.0, indicating excellent detection of positive samples, albeit with slightly lower specificity (82.4% and 81.1%, respectively), which may reflect an increase in false positives. Hospital C exhibited a more balanced trade-off between specificity was 0.889 and sensitivity 0.887. Overall, the model demonstrated strong adaptability and robustness across different hospitals, maintaining high performance despite potential variations in data distribution.

Table 3
The Performance of TGMIL on val, and external test sets. Center¹, Center², Center³, and Center⁴ are Shanxi Provincial People's Hospital, Tenth People's Hospital of the Tongji University, Shantou Central Hospital, and Xiang'an Hospital of Xiamen University, respectively.

		AUC	Accuracy	Specificity	Sensitivity
Val set		0.772 ± 0.084	0.831 ± 0.047	0.887 ± 0.043	0.906 ± 0.040
Test set	Center1	0.837 ± 0.053	0.828 ± 0.034	0.824 ± 0.034	1.000 ± 0.000
	Center2	0.886 ± 0.072	0.828 ± 0.081	0.911 ± 0.085	0.769 ± 0.122
	Center3	0.814 ± 0.091	0.859 ± 0.065	0.889 ± 0.077	0.887 ± 0.074
	Center4	0.812 ± 0.156	0.846 ± 0.085	0.811 ± 0.099	1.000 ± 0.000

3.4 Diagnostic performance of the model alone, the ultrasonographers alone, and the model-assisted ultrasonographers

We compared the diagnostic performance across ultrasonographers with different levels experience. As anticipated, attending ultrasonographers performed better than junior ultrasonographers, and senior ultrasonographers exhibited the highest performance among the three groups. The sensitivity, specificity, accuracy, and AUC for attending vs junior ultrasonographers were 80.8% vs 69.2%, 80.7% vs 71.2%, 81.8% vs 69.2%, and 0.808 [95% CI: 0.735–0.880] vs 0.698 [95% CI: 0.615–0.780] (P < 0.05). Additionally, the sensitivity, specificity, accuracy, and AUC for the senior vs. attending

ultrasonographers were 91.1% vs 80.8%, 88.5% vs 80.77%, 89.9% vs 81.8%, and 0.898 [95% CI: 0.841–0.9555] vs 0.807 [95% CI: 0.735–0.8880] ($P < 0.05$).

We also compared the diagnostic performance of ultrasonographers without or with TGMIL assistance. The pooled sensitivity for all ultrasonographers without TGMIL assistance was 82.7%, which improved to 86.5% ($P < 0.05$) with TGMIL assistance. For junior ultrasonographers, sensitivity increased from 71.2% to 82.7% ($P < 0.05$) with TGMIL assistance, and for attending ultrasonographers, sensitivity improved from 80.7% to 86.5% ($P < 0.05$) with TGMIL assistance. The accuracy of all ultrasonographers without TGMIL assistance was 80.8%, which was improved to 87.9% ($P < 0.05$) with TGMIL assistance. For the junior ultrasonographers, the accuracy was improved from 69.2% to 85.9% ($P < 0.05$) and for the attending ultrasonographers the accuracy was improved from 81.8% to 87.3% ($P < 0.05$) with TGMIL assistance. Specificity was significantly improved for junior ultrasonographers, from 71.2% to 82.7% ($P < 0.05$) with TGMIL assistance. The AUC of all ultrasonographers without TGMIL assistance was 0.814 [95% CI: 0.744–0.885], which was improved to 0.875 [95% CI: 0.813–0.936] ($P < 0.05$) with TGMIL assistance. For the junior ultrasonographers, the AUC was improved from 0.698 [95% CI: 0.615–0.781] to 0.845 [95% CI: 0.778–0.913] ($P < 0.05$), and for the attending ultrasonographers the AUC was improved from 0.808 [95% CI: 0.735–0.880] to 0.871 [95% CI: 0.809–0.933] ($P < 0.05$) with TGMIL assistance. For senior ultrasonographers, the diagnostic performance was not significantly improved with the model in any diagnostic index, including sensitivity, specificity, accuracy, or AUC. All results are shown in Table 4.

Table 4

Comparison of the diagnostic performance of TGMIL and ultrasonographers of different experience levels (test set).

TGMIL	AUC (95% CI)	Accuracy (%)	Sensitivity(%)	Specificity (%)	PPV(%)	NPV(%)	Kappa
	0.824 (0.755–0.894)	91.9(182/198)	93.8 (137/146)	86.5 (45/52)	95.1 (137/144)	83.3 (45/54)	
Ultrasonographers without DMAAN assistance							
All	0.814 [†] (0.744–0.885)	80.8 [†] (160/198)	80.1 [†] (117/146)	82.7 (43/52)	92.8 (117/126)	59.7 (43/72)	0.559
Junior	0.698 [†] (0.615–0.781)	69.2 [†] (137/198)	68.4 [†] (100/146)	71.2 [†] (37/52)	87.0 (100/115)	44.6 (37/83)	0.325
Attending	0.808 ^{†‡} (0.735–0.880)	81.8 ^{†‡} (160/198)	80.8 ^{†‡} (118/146)	80.7 [†] (42/52)	92.2(118/128)	66.0 (42/70)	0.548
Senior	0.898 [§] (0.841–0.955)	89.9 [§] (179/198)	91.1 ^{†§} (133/146)	88.5 [§] (46/52)	95.6 (133/139)	77.9 (46/59)	0.760
Ultrasonographers with DMAAN assistance							
All	0.875 [¶] (0.813–0.936)	87.9 [¶] (174/198)	88.4 [¶] (129/146)	86.5(45/52)	94.8(129/136)	72.6(45/62)	0.701
Junior	0.845 [¶] (0.778–0.913)	85.9 [¶] (170/198)	86.3 [¶] (126/146)	82.7 [¶] (43/52)	94.0(126/134)	67.2(43/64)	0.651
Attending	0.871 [¶] (0.809–0.933)	87.3 [¶] (173/198)	87.7 [¶] (128/146)	86.5(45/52)	94.8(128/135)	71.4(45/63)	0.694
Senior	0.885(0.824–0.945)	89.44 (177/198)	90.4(132/146)	86.5(45/52)	95.0(132/139)	76.3(45/59)	0.742
AUC: area under the receiver operating characteristic curve; NPV: negative predictive value; PPV: positive predictive value; CI: confidence interval.							
†: $P < 0.05$ for comparison of diagnostic performance of TGMIL-assisted process and TGMIL alone; ¶: $P < 0.05$ for comparison of diagnostic performance without TGMIL; ‡: $P < 0.05$ for comparison of diagnostic performance with junior ultrasonographers; §: $P < 0.05$ for comparison of diagnostic performance with attending ultrasonographers.							

4. Discussion

In this study, we developed a DL system (TGMIL) for ultrasound video detection based on global and local aggregation. This model effectively aggregate long-range semantic information to alleviate the problem of category ambiguity, and the aggregation of adjacent frame motion information can effectively assist in accurately localization. Importantly, the newly developed DL model can not only automatically identify the breast lesions with BI-RADS category 4–5, but can also improve the diagnostic efficiency of less-experienced (< 10 years) ultrasonographers.

To explore the diagnostic performance of the DL network in breast ultrasound videos, we compared dynamic and traditional static image classification models as supplementary tools for assessing benign and malignant breast lesions. As reported in

previous studies, DL ultrasound models with traditional image classification models have outperformed healthcare professionals with respect to certain clinical tasks ^[34]. In clinical practice, video ultrasound is capable of showing the shape, boundaries, internal structure, and calcifications of the lesion more comprehensively. Thus, the dynamic videos permit real-time diagnosis ^[28], and offer more thorough information in terms of differential diagnosis of breast lesions.

we also investigated the value of TGMIL model in supporting and navigating ultrasonographers to improve diagnostic performance. TGMIL increased both negative and positive predictive values (68.4% and 90.2%, respectively). The model demonstrated effective diagnostic capabilities on annotated ultrasound video sequences, which is comparable to the level of senior physicians. Our model demonstrates the potential benefits in automatic identification of lesions and improving diagnostic accuracy, particularly for junior and attending ultrasonographers. In addition, TGMIL may help avoid unnecessary invasive biopsies through increasing negative predictive value. In the future, this model could be applied for screening and early detection as new directions.

However, this study has several limitations. The cases included in the training and test sets were confirmed through pathological examinations conducted in a hospital setting and included only BI-RADS 4–5 lesions, excluding other categories. Hence, the results may significantly affect the positive and negative predictive values of only a subset of the population, and limiting the potential generalizability of the findings. Moreover, the accuracy and stability of TGMIL models depend on the gold standard of cyto- and histo-pathological diagnosis, which is difficult to obtain for the general population at routine testing stations. Our primary objective was to evaluate the diagnostic accuracy of dynamic ultrasound videos when the TGMIL model was used to improve efficiency. Therefore, caution is warranted when extrapolating these findings to other contexts. However, in the clinical setting, TGMIL may enhance the performance of ultrasonographers after they review ultrasound videos and traditional image classification models. Our results suggest that ultrasonographers with TGMIL performed better than those in a previous study ^[35].

Declarations

Acknowledgements

Not applicable.

Funding

This study was funded by the National Natural Science Foundation of China (No. 32171363), the Key Research and Development Program for Social Development of Yunnan Science and Technology Department (202403AC100014), International Joint Laboratory on Cancer Molecular Imaging of Yunnan (202503AP140032), Jie Tian Specialist Workstation of Yunnan (202505AF350067), Research Project of Yunnan Clinical Medical Center (2024YNLCYXZX0418) and horizontal project funding from the third Affiliated Hospital of Kunming Medical University (20233160A0866, 20243160A0511). Additionally, it received funding from the start-up fund for Fujian Key Laboratory of Precision Diagnosis and Treatment in Breast Cancer for Guojun Zhang, as well as start-up and supporting Funds from the Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital for Guojun Zhang..

Author contributions

W.B.L., Z.C.D., W.H.H., L.S.W., and G.J.Z., participated in conceptualization, methodology, resources, writing of the original draft, supervision and funding acquisition. R.Y.Q. participated in the conceptualization, methodology, formal data analysis and writing of the original draft. Y.C.C., L.S.L., Z.J.Y., L.H.G., and M.Q.H., participated in the conceptualization, methodology and writing of the original draft. All authors contributed to the research, editing and approval of the paper.

Ethics declarations

Ethics approval and consent to participate by Ethics Committee of Xiang'an Hospital of Xiamen University (XAHLL2022060)

Consent for publication

Not applicable.

Competing interest

The authors declare that they have no conflicts of interest.

Data availability

The main data supporting the results in this study are available within the article and its Supplementary Information. Individual-level patient data are protected because of patient privacy; they are accessible with the consent of the data management committee from institutions and are not publicly available. The data management committee will then review all the requests and grant permission (if successful). A formal data transfer agreement will be required upon approval. Generally, all these requests for access to the data will be responded to within 1 month. All data shared will be deidentified. Source data are provided with this paper.

Additional information

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Kulaçoğlu S, Erkinliç G. Imp3 expression in benign and malignant thyroid tumors and hyperplastic nodules [J]. *Balkan Med J*, 2015, 32(1): 30–7.
2. Berg W A, Campassi C, Langenberg P, et al. Breast Imaging Reporting and Data System [J]. 2012.
3. Magny S J, Shikhman R, Keppke A L. Breast Imaging Reporting and Data System [M]. StatPearls. Treasure Island (FL); StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC. 2022.
4. Ohuchi N, Suzuki A, Sobue T, et al. Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START): a randomised controlled trial [J]. *Lancet*, 2016, 387(10016): 341–8.
5. Bae J M, Shin S Y, Kim E H, et al. Distribution of dense breasts using screening mammography in Korean women: a retrospective observational study [J]. *Epidemiol Health*, 2014, 36: e2014027.
6. Ding R, Xiao Y, Mo M, et al. Breast cancer screening and early diagnosis in Chinese women [J]. *Cancer Biol Med*, 2022, 19(4): 450–67.
7. Breast Cancer Expert Committee of National Cancer Quality Control Center; Breast Cancer Expert Committee of China Anti-Cancer Association; Cancer Drug Clinical Research Committee of China Anti-Cancer Association. [Guidelines for diagnosis and treatment of advanced breast cancer in China (2024 edition)]. *Zhonghua Zhong Liu Za Zhi*. 2024;46(12):1079–1106.
8. Maskarinec G, Pagano I, Chen Z, et al. Ethnic and geographic differences in mammographic density and their association with breast cancer incidence [J]. *Breast Cancer Res Treat*, 2007, 104(1): 47–56.
9. Merritt, Rb C. Combined Screening With Ultrasound and Mammography vs Mammography Alone in Women at Elevated Risk of Breast Cancer [J]. *Journal of the American Medical Association*, 2008, 299(4): 2151–63.
10. Berg W A, Blume J D, Cormack J B, et al. Operator Dependence of Physician-performed Whole-Breast US: Lesion Detection and Characterization1 [J]. *Radiology*, 2006, 241(2): 355–65.
11. Merry G M, Mendelson E B. Update on Screening Breast Ultrasonography [J]. *Radiologic Clinics of North America*, 2014, 52(3): 527–37.

12. Cheng X, Li H, Chen Y, et al. Ultrasound-triggered phase transition sensitive magnetic fluorescent nanodroplets as a multimodal imaging contrast agent in rat and mouse model [J]. *PLoS One*, 2013, 8(12): e85003.
13. Corsetti V, Houssami N, Ferrari A, et al. Breast screening with ultrasound in women with mammography-negative dense breasts: evidence on incremental cancer detection and false positives, and associated cost [J]. *Eur J Cancer*, 2008, 44(4): 539–44.
14. Lecun Y, Bengio Y, Hinton G. Deep learning [J]. *Nature*, 2015, 521(7553): 436.
15. Lyu S Y, Zhang Y, Zhang M W, et al. Diagnostic value of artificial intelligence automatic detection systems for breast BI-RADS 4 nodules [J]. 2022, 10(2): 10.
16. Krishna S, Ss S, Krishnamoorthy S, et al. Stain Normalized Breast Histopathology Image Recognition using Convolutional Neural Networks for Cancer Detection [J]. 2022.
17. Deng J, She Y, Wang J, et al. Invasiveness assessment of deep learning method for pulmonary subsolid nodules; proceedings of the ERS International Congress 2020 abstracts, F, 2020 [C].
18. Lee S E, Moon J E, Rho Y H, et al. Which supplementary imaging modality should be used for breast ultrasonography? Comparison of the diagnostic performance of elastography and computer-aided diagnosis [J]. *Ultrasonography*, 2017, 36(2): 153–9.
19. Jeongmin, Lee, Sanghee, et al. Evaluation of the effect of computer aided diagnosis system on breast ultrasound for inexperienced radiologists in describing and determining breast lesions [J]. *Medical ultrasonography*, 21(3): 239–45.
20. Liang Y, Ping Z, Zhang J, et al. Performance of Computer-Aided Diagnosis in Ultrasonography for Detection of Breast Lesions Less and More Than 2 cm: Prospective Comparative Study [J]. *JMIR Medical Informatics*, 2020, 8(3).
21. Park H J, Kim S M, Yun B L, et al. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound: Added value for the inexperienced breast radiologist [J]. *Medicine*, 2019, 98(3).
22. Romeo V, Cuocolo R, Apolito R, et al. Clinical value of radiomics and machine learning in breast ultrasound: a multicenter study for differential diagnosis of benign and malignant lesions [J]. *Eur Radiol*, 2021, 31(12): 9511–9.
23. Pfof A, Sidey-Gibbons C, Barr R G, et al. Intelligent multi-modal shear wave elastography to reduce unnecessary biopsies in breast cancer diagnosis (INSPIRED 002): a retrospective, international, multicentre analysis [J]. *Eur J Cancer*, 2022, 177: 1–14.
24. Yu T F, He W, Gan C G, et al. Deep learning applied to two-dimensional color Doppler flow imaging ultrasound images significantly improves diagnostic performance in the classification of breast masses: a multicenter study [J]. *Chin Med J (Engl)*, 2021, 134(4): 415–24.
25. Ma M, Mei S, Wan S, et al. Keyframe Extraction From Laparoscopic Videos via Diverse and Weighted Dictionary Selection [J]. *IEEE J Biomed Health Inform*, 2021, 25(5): 1686–98.
26. Huang R, Ying Q, Lin Z, et al. Extracting keyframes of breast ultrasound video using deep reinforcement learning [J]. *Med Image Anal*, 2022, 80: 102490.
27. Zhang X, Liang M, Yang Z, et al. Deep Learning-Based Radiomics of B-Mode Ultrasonography and Shear-Wave Elastography: Improved Performance in Breast Mass Classification [J]. *Front Oncol*, 2020, 10: 1621.
28. Fujioka T, Mori M, Kubota K, et al. Breast Ultrasound Image Synthesis using Deep Convolutional Generative Adversarial Networks [J]. *Diagnostics (Basel)*, 2019, 9(4).
29. Tanaka H, Chiu S W, Watanabe T, et al. Computer-aided diagnosis system for breast ultrasound images using deep learning [J]. *Phys Med Biol*, 2019, 64(23): 235013.
30. Wang J, Jiang J, Zhang D, et al. An integrated AI model to improve diagnostic accuracy of ultrasound and output known risk features in suspicious thyroid nodules [J]. *Eur Radiol*, 2022, 32(3): 2120–9.
31. Lee J H, Novak C L, Aylward S, et al. Computer-aided lesion diagnosis in B-mode ultrasound by border irregularity and multiple sonographic features [J]. 2013, 8670: 867010.

32. Guo R, Lu G, Qin B, et al. Ultrasound Imaging Technologies for Breast Cancer Detection and Management: A Review [J]. Ultrasound Med Biol, 2018, 44(1): 37–70.

33. Nasiri-Sarvi A, Hosseini M S, Rivaz H. Vision Mamba for Classification of Breast Ultrasound Images [J]. 2024.

34. Han S, Kang H K, Jeong J Y, et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images [J]. Phys Med Biol, 2017, 62(19): 7714–28.

35. Gao Y, Liu B, Zhu Y, et al. Detection and recognition of ultrasound breast nodules based on semi-supervised deep learning: a powerful alternative strategy [J]. Quant Imaging Med Surg, 2021, 11(6): 2265–78. strategy [J]. Quant Imaging Med Surg, 2021, 11(6): 2265–78.

Figures

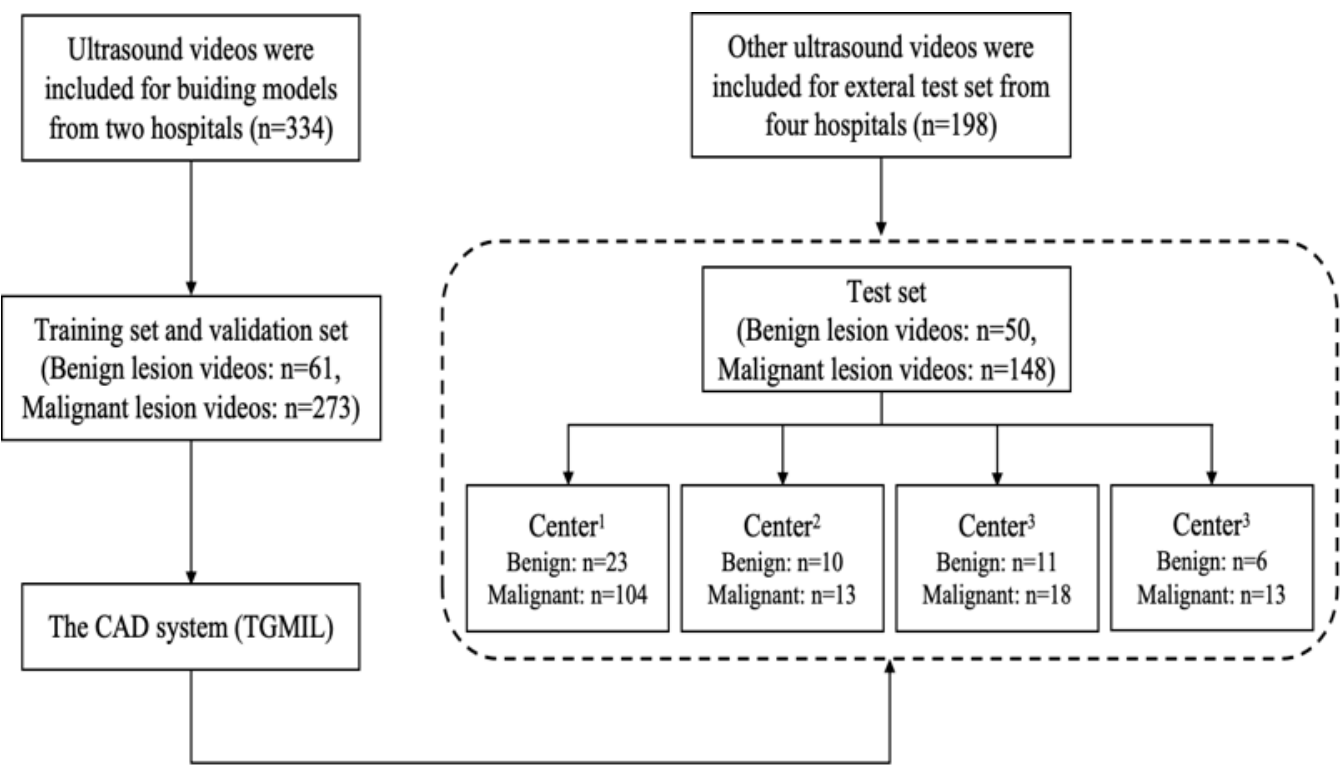


Figure 1

Legend not included with this version

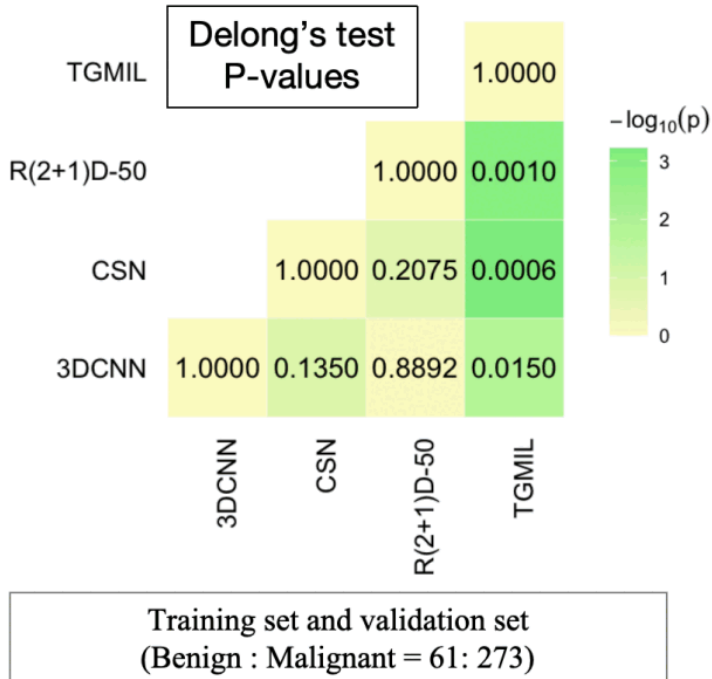
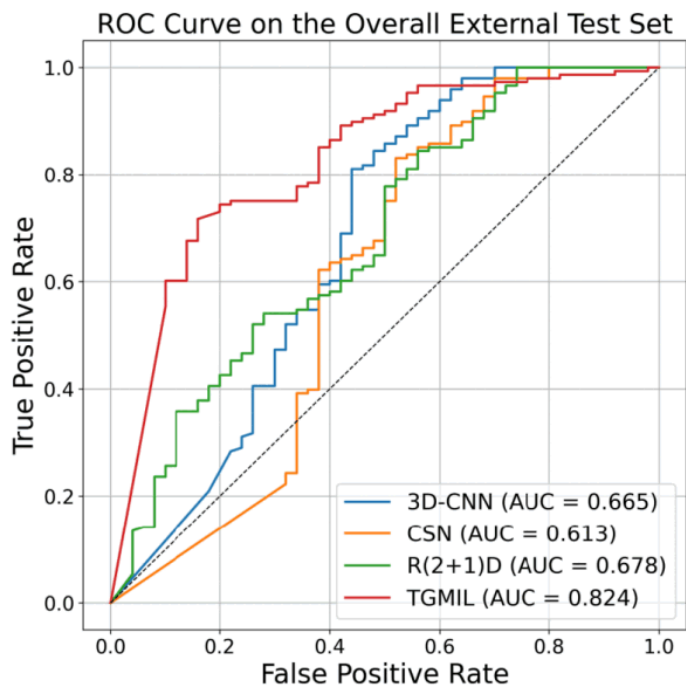


Figure 2

Legend not included with this version

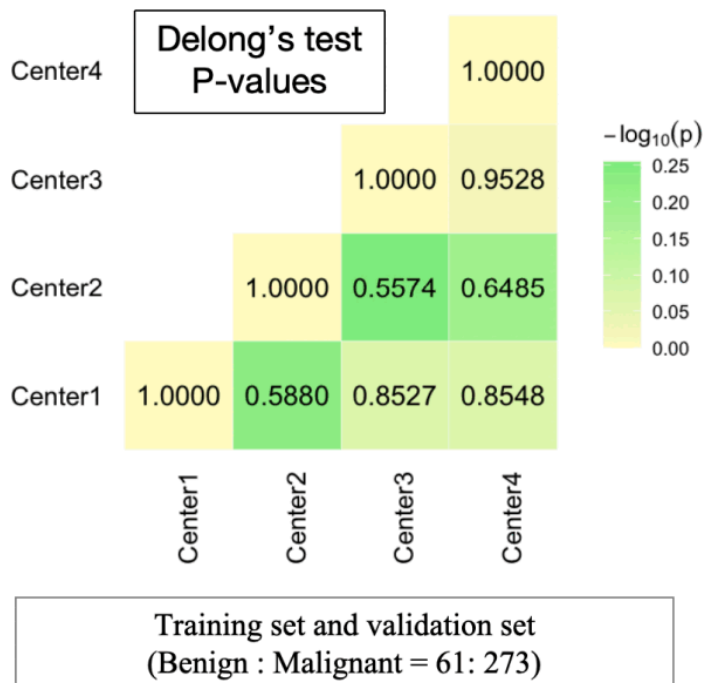
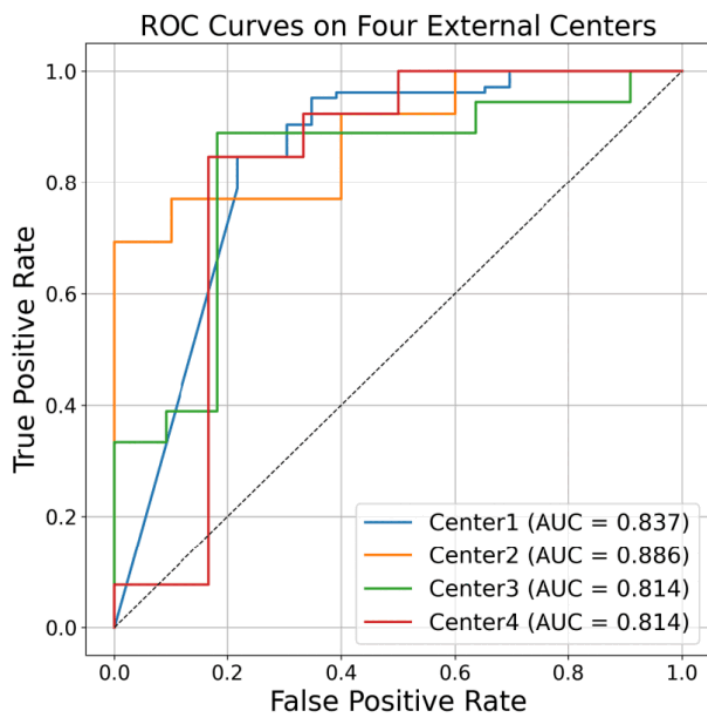


Figure 3

Legend not included with this version

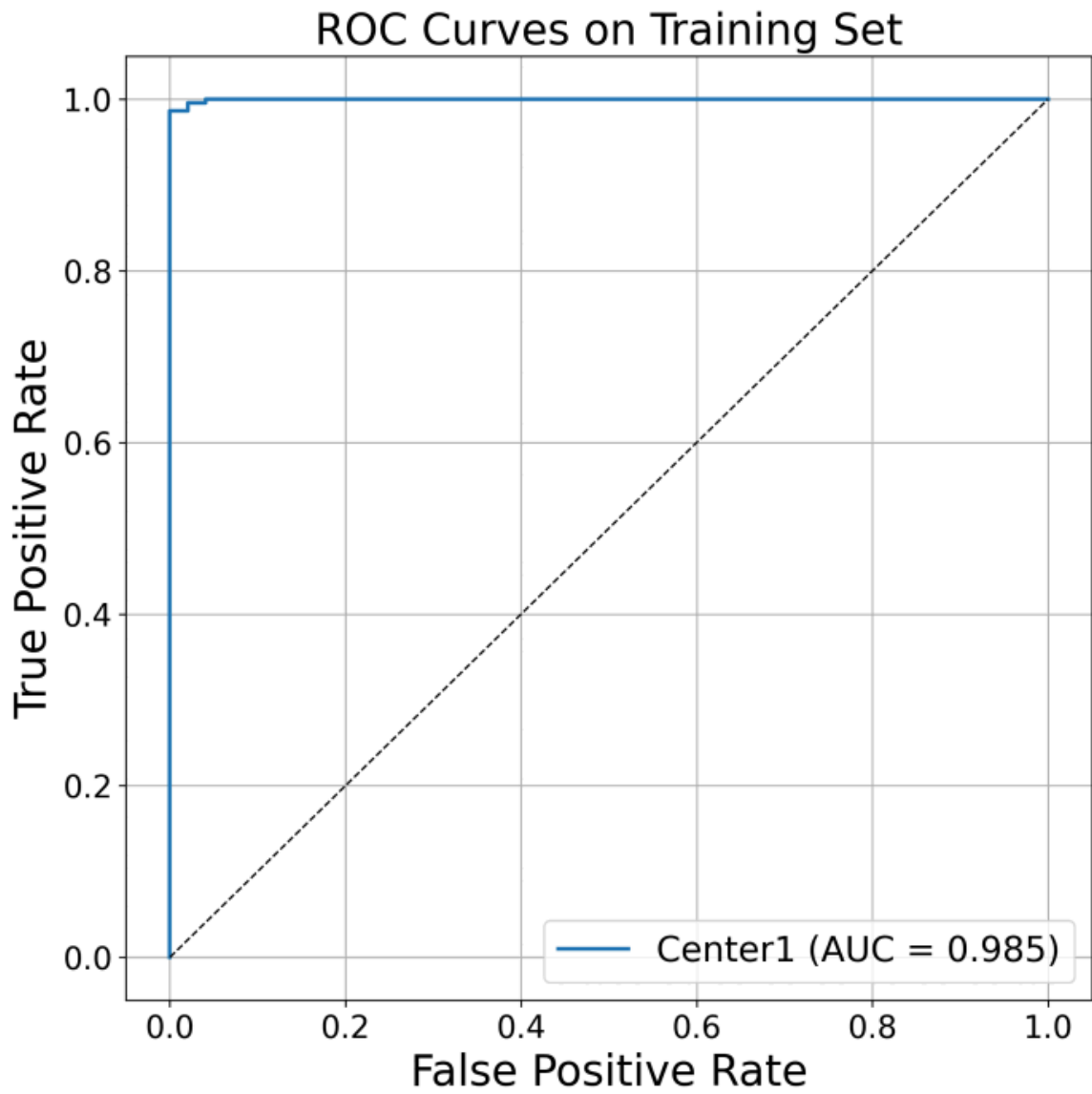


Figure 4

Legend not included with this version