biomedicines

MDPI

Article

Assessing Detection Accuracy of Computerized Sonographic Features and Computer-Assisted Reading Performance in Differentiating Thyroid Cancers

Hao-Chih Tai [1], Kuen-Yuan Chen [1], Ming-Hsun Wu [1], King-Jen Chang [1], Chiung-Nien Chen [1,*] and Argon Chen [2,*]

[1] Department of Surgery, National Taiwan University Hospital and College of Medicine, Taipei 100225, Taiwan; taihc1@ntu.edu.tw (H.-C.T.); dtsurg51@ntuh.gov.tw (K.-Y.C.); 010028@ntuh.gov.tw (M.-H.W.); kingjen@ntu.edu.tw (K.-J.C.)

[2] Graduate Institute of Industrial Engineering, National Taiwan University, Taipei 106216, Taiwan

* Correspondence: cnchen@ntu.edu.tw (C.-N.C.); achen@ntu.edu.tw (A.C.)

Abstract: For ultrasound imaging of thyroid nodules, medical guidelines are all based on findings of sonographic features to provide clinicians management recommendations. Due to the recent development of artificial intelligence and machine learning (AI/ML) technologies, there have been computer-assisted detection (CAD) software devices available for clinical use to detect and quantify the sonographic features of thyroid nodules. This study is to validate the accuracy of the computerized sonographic features (CSF) by a CAD software device, namely, AmCAD-UT, and then to assess how the reading performance of clinicians (readers) can be improved providing the computerized features. The feature detection accuracy is tested against the ground truth established by a panel of thyroid specialists and a multiple-reader multiple-case (MRMC) study is performed to assess the sequential reading performance with the assistance of the CSF. Five computerized features, including anechoic area, hyperechoic foci, hypoechoic pattern, heterogeneous texture, and indistinct margin, were tested, with AUCs ranging from 0.888~0.946, 0.825~0.913, 0.812~0.847, 0.627~0.77, and 0.676~0.766, respectively. With the five CSFs, the sequential reading performance of 18 clinicians is found significantly improved, with the AUC increasing from 0.720 without CSF to 0.776 with CSF. Our studies show that the computerized features are consistent with the clinicians' findings and provide additional value in assisting sonographic diagnosis.

Keywords: thyroid nodules; sonographic features; computer-assisted detection; MRMC study

check for updates

Citation: Tai, H.-C.; Chen, K.-Y.; Wu, M.-H.; Chang, K.-J.; Chen, C.-N.; Chen, A. Assessing Detection Accuracy of Computerized Sonographic Features and Computer-Assisted Reading Performance in Differentiating Thyroid Cancers. *Biomedicines* **2022**, *10*, 1513. https://doi.org/10.3390/biomedicines10071513

Academic Editor: Su Woong Yoo

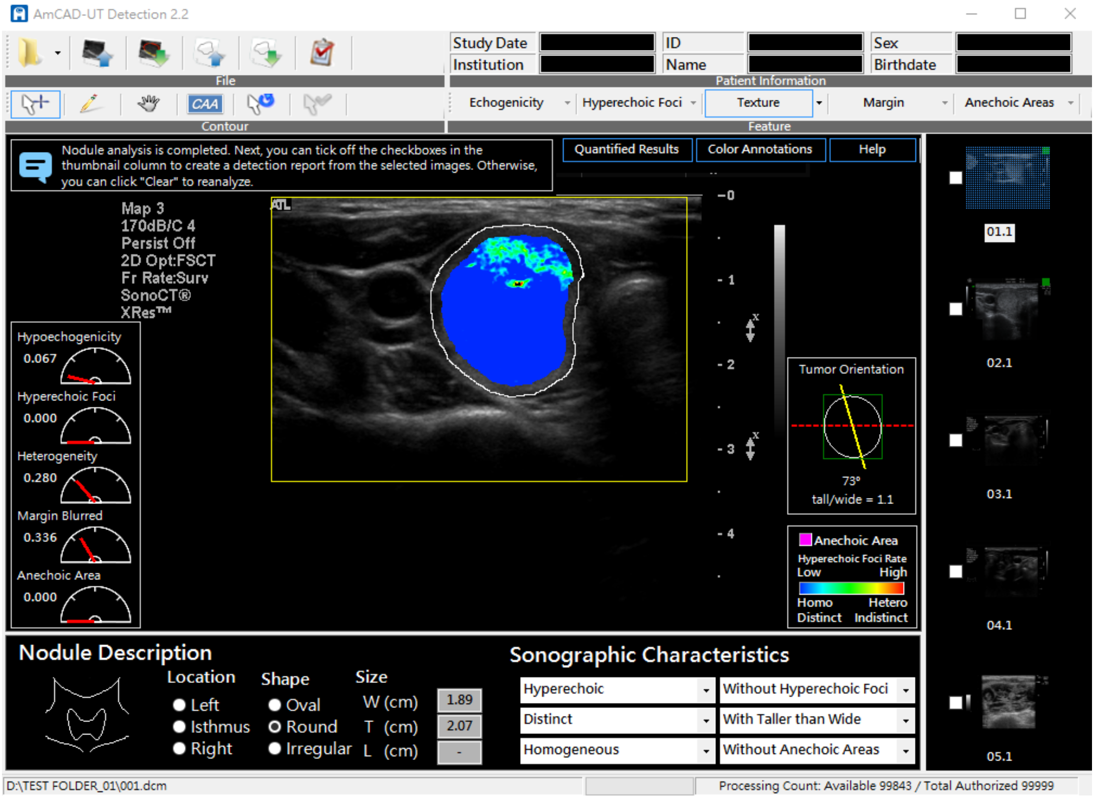
Received: 30 May 2022
Accepted: 24 June 2022
Published: 26 June 2022

1. Introduction

Thyroid cancer is the most common endocrine cancer, and its incidence has increased dramatically by an average of 4.5% annually [1]. Accurate identification of thyroid cancer is crucial for effective treatment. Ultrasonography is the most common tool for early detection of thyroid cancer because it is readily available and noninvasive. In the past decade, use of high-resolution ultrasound has resulted in improved detection of thyroid nodules [2,3]. Nevertheless, most of the nodules are benign, and thyroid cancers only account for 7–15% of detected nodules [4]. Identification of malignant nodules is critical to avoid unnecessary fine-needle aspiration (FNA) biopsy and surgical procedures. Medical guidelines, most notably the management guidelines by the American Thyroid Association (ATA) [4] and Thyroid Imaging Reporting and Data System (TI-RADS) by the American College of Radiology (ACR) [5], have been developed and recommended to clinicians on how to identify and use the sonographic features for differentiation of thyroid cancers. Important features include micro-calcifications, hypo-echogenicity, irregular margins, taller-than-wide shape, etc. [4,6–10]. However, presence of the sonographic features is determined based on a physician's subjective interpretation, which may be influenced by education and experience. Interpretation discrepancies among clinicians or even by the same clinicians at different times have become a major issue that hinders the diagnosis and treatment of thyroid nodules [11,12].

Recent development of AI/ML technologies has given rise to the use of CAD software devices in clinical practice [13,14], to assist clinicians with improving their diagnosis accuracy and workflow effectiveness. The CAD solutions have been applied to ultrasound imaging of various diseases, such as breast cancer [15,16] and liver lesions [17]. One CAD software device has emerged to give a second opinion on image interpretation [18], to reduce inter-observer variation in breast images [16,19]. Studies have also reported applications of CAD devices to ultrasound imaging of thyroid nodules [20–28]. In particular, the effectiveness of computerized sonographic features (CSF) to differentiate malignant nodules has been demonstrated [20–24]. The CSFs provided by a CAD device are used to assist clinicians in interpreting the images and then making recommendations based on the clinicians' professional judgment and/or medical guidelines. Although both clinicians' finding and computerization of sonographic features are shown helpful in thyroid cancer diagnosis, the computerized sonographic features have not yet been validated by comparing to the clinicians' sonographic feature findings and by studying their effect in assisting clinicians' reading of the thyroid sonograms.

In this study, an FDA-cleared CAD software device, AmCAD-UT (AmCad BioMed Co., Taipei, Taiwan) (Figure 1a), is employed to validate the software's detection and quantification of the sonographic features (Figure 1b) against the ground truth established by a panel of thyroid specialists and then to perform an MRMC study to test the reader's performance sequentially assisted with the computerized features calculated by the CAD device.



AmCAD-UT Detection 2.2

Study Date [redacted] ID [redacted] Sex [redacted]
Institution [redacted] Name [redacted] Birthdate [redacted]

Nodule analysis is completed. Next, you can tick off the checkboxes in the thumbnail column to create a detection report from the selected images. Otherwise, you can click "Clear" to reanalyze.

Map 3
170dB/C 4
Persist Off
2D Opt:FSCT
Fr Rate:Surv
SonoCT®
XRes™

Hypoechogenicity 0.067
Hyperechoic Foci 0.000
Heterogeneity 0.280
Margin Blurred 0.336
Anechoic Area 0.000

Tumor Orientation
73°
tall/wide = 1.1

Nodule Description

Location: ● Left ● Isthmus ● Right
Shape: ● Oval O Round ● Irregular
Size: W (cm) 1.89; T (cm) 2.07; L (cm) -

Sonographic Characteristics

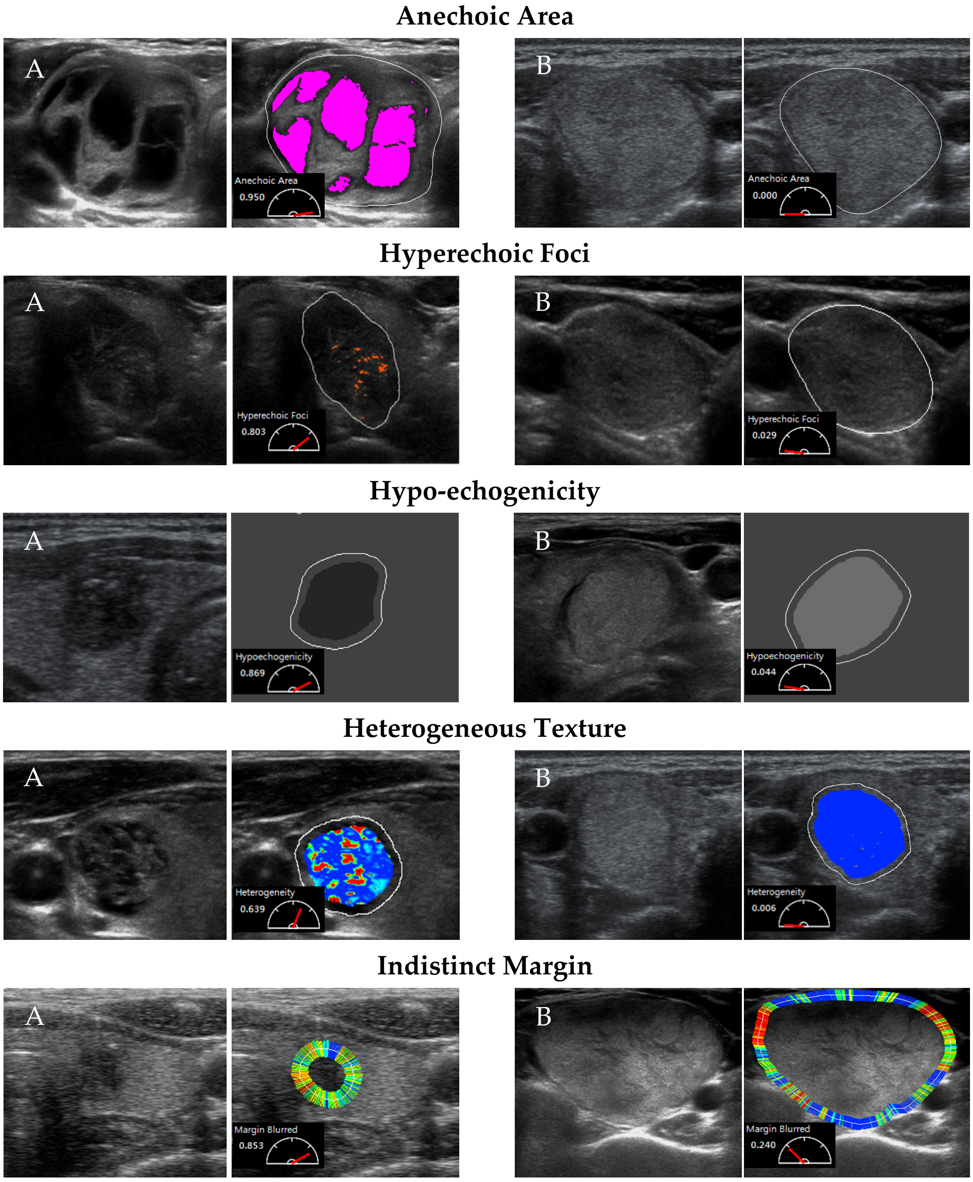
Hyperechoic | Without Hyperechoic Foci
Distinct | With Taller than Wide
Homogeneous | Without Anechoic Areas

D:\TEST FOLDER_01\001.dcm

Processing Count: Available 99843 / Total Authorized 99999

(**a**)

Figure 1. *Cont.*



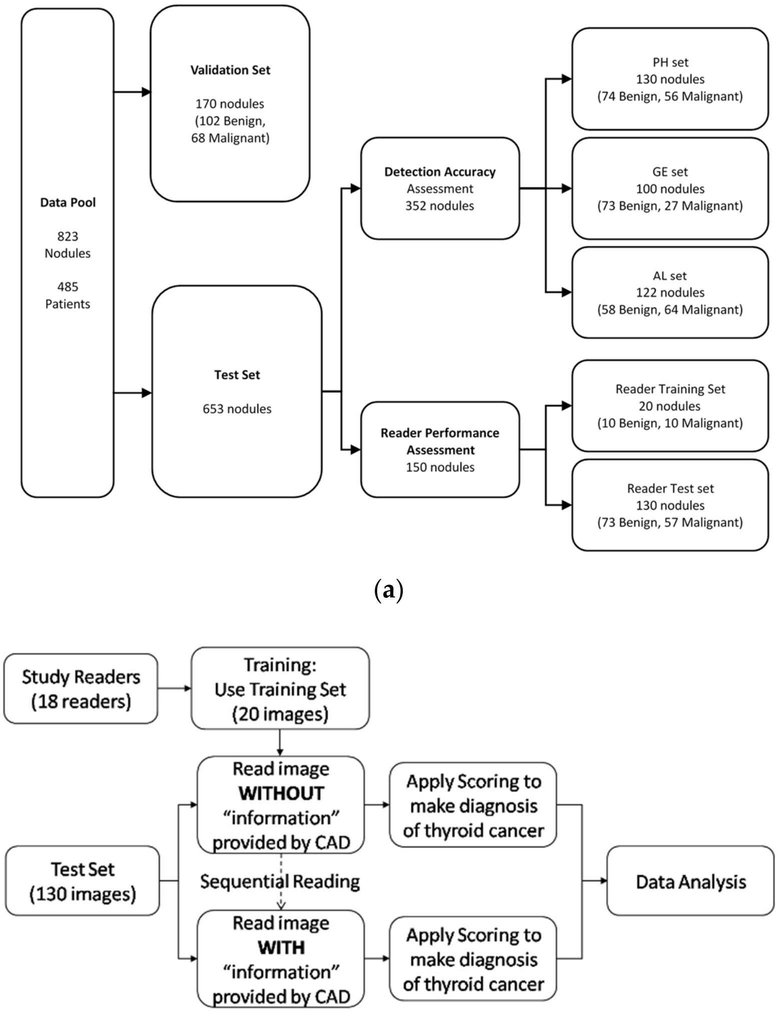
(**b**)

Figure 1. AmCAD-UT: (**a**) Graphical user interface; (**b**) A: nodules with features present; B: nodules without the features. Features are visualized, with colors indicating the severity or likelihood of the feature presence—the warmer the color the more likely or severer the feature's presence. Quantified values of the computerized features ranging from 0 to 1 are shown and displayed in the pointer meters.

2. Materials and Methods

2.1. Database for Computerized Features Testing

The Institutional Review Board of the National Taiwan University Hospital (NTUH) approved this prospective study (200805039R). Informed consent was obtained from all participants and all patient identifiers were removed from the images used in the study. The database consisted of a collection of thyroid sonograms of patients who underwent a thyroidectomy due to suspicious thyroid carcinoma, follicular neoplasm, or symptomatic nodular goiter diagnosed by ultrasound imaging and FNA cytology, at NTUH. Since the quality of the sonograms greatly depended on the ultrasound scanners and might affect both the clinicians' finding and computerization of the sonographic features, we collected for this study sonograms obtained from different ultrasound scanners. The sonograms were acquired in DICOM format using Philips HDI 5000 (denoted as PH), GE Voluson 730 PRO (denoted as GE), and ALOKA Prosound2 (denoted as AL) ultrasound scanners, with 5–12 MHz linear multifrequency probes under identical imaging setting by certified technicians. The images were ineligible if the nodule size was larger than the width of the probe array or if the single nodule was not separable from another in cases of multinodular goiters. Finally, sonograms of 823 nodules (663 patients) were included in the database (Figure 2a). The major ethnic group was Chinese. Our previous studies investigated the effectiveness of specific computerized sonographic features, namely, calcification, heterogeneity, or echogenicity, in distinguishing malignant from benign nodules [20–24], whereas the current study validate the detection accuracy against the ground truth established by a panel of three thyroid specialists. A set of images was used for algorithm validation, consisting of 170 sonograms (102 benign and 68 malignant). In total, 653 sonograms were used for testing. Of these, 352 sonograms were randomly chosen for test of feature detection accuracy, and another 150 images from the Philips scanner were chosen for reading performance assessment (Table 1a). The pathology results of the nodules and cancer type distribution are shown in Table 1b. Among the 150 sonograms, 20 (10 benign and 10 malignant) were used as the training set for the 18 readers to get familiar with the CAD software user interface and to train themselves using the software to differentiate malignant nodules. The remaining 130 sonograms were used for the MRMC study (Figure 2b).


Data Pool
823 Nodules
485 Patients
Validation Set
170 nodules
(102 Benign, 68 Malignant)
Test Set
653 nodules
Detection Accuracy Assessment
352 nodules
Reader Performance Assessment
150 nodules
PH set
130 nodules
(74 Benign, 56 Malignant)
GE set
100 nodules
(73 Benign, 27 Malignant)
AL set
122 nodules
(58 Benign, 64 Malignant)
Reader Training Set
20 nodules
(10 Benign, 10 Malignant)
Reader Test set
130 nodules
(73 Benign, 57 Malignant)
Study Readers (18 readers)
Training: Use Training Set (20 images)
Test Set (130 images)
Read image WITHOUT "information" provided by CAD
Sequential Reading
Read image WITH "information" provided by CAD
Apply Scoring to make diagnosis of thyroid cancer
Apply Scoring to make diagnosis of thyroid cancer
Data Analysis

(**a**)

(**b**)

Figure 2. Study datasets: (**a**) Datasets for the validation, detection accuracy test, and reader performance assessment. (**b**) Datasets and readers for the MRMC reader performance study.

Table 1. Demographics of the patients and nodules.

	Detection Accuracy Assessment **PH Set**			**Detection Accuracy Assessment** **GE Set**		
	Patients (*n* = 113)			**Patients (*n* = 84)**		
	Benign **(*n* = 74)**	**Malignant** **(*n* = 56)**	**Total** **(*n* = 130)**	**Benign** **(*n* = 73)**	**Malignant** **(*n* = 27)**	**Total** **(*n* = 100)**
Gender No. (%)						
Male	14 (18.9)	9 (16.1)	23 (17.7)	14 (19.2)	3 (11.1)	17 (17.0)
Female	60 (81.1)	47 (83.9)	107 (82.3)	59 (80.8)	24 (88.9)	83 (83.0)
Age (year)						
Mean ± SD *	48.2 ± 13.8	48.6 ± 15.4	48.3 ± 14.5	51.1 ± 13.2	50.1 ± 16.6	50.5 ± 14.1
Range	20.9–76.9	11.2–71.6	11.2–76.9	20.9–75.3	20.4–84.8	20.4–84.8
Size of Nodule (cm)						
Mean ± SD *	2.38 ± 0.88	1.87 ± 0.86	2.16 ± 0.90	2.50 ± 0.94	1.86 ± 0.72	2.33 ± 0.93
Range	0.94–4.37	0.49–4.09	0.49–4.37	0.71–4.16	0.87–3.66	0.71–4.16

(**a**) Demographics of patients and nodule images acquired from Philips (PH), GE, and Aloka (AL) ultrasound scanners for detection accuracy and reader performance assessment.

	Detection Accuracy Assessment **AL Set**			**Reader** **Performance Assessment**		
	Patients (*n* = 107)			**Patients (*n* = 129)**		
	Benign **(*n* = 58)**	**Malignant** **(*n* = 64)**	**Total** **(*n* = 122)**	**Benign** **(*n* = 83)**	**Malignant** **(*n* = 67)**	**Total** **(*n* = 150)**
Gender No. (%)						
Male	8 (13.8)	14 (21.9)	22 (18.0)	16 (19.3)	11 (16.4)	27 (18.0)
Female	50 (86.2)	50 (78.1)	100 (82.0)	67 (80.7)	56 (83.6)	123 (82.0)
Age (year)						
Mean ± SD *	45.9 ± 9.5	42.7 ± 11.8	44.2 ± 10.8	48.5 ± 13.9	47.8 ± 14.8	48.2 ± 14.3
Range	21.7–64.3	1.5–74.0	1.5–74.0	20.9–76.9	11.2–71.6	11.2–76.9
Size of Nodule (cm)						
Mean ± SD *	2.45 ± 1.06	1.48 ± 0.72	1.94 ± 1.02	2.38 ± 0.86	1.86 ± 0.83	2.15 ± 0.88
Range	0.53–4.29	0.53–4.12	0.53–4.29	0.94–4.37	0.49–4.09	0.49–4.37

(**b**) The pathology diagnosis results of the 823 nodules.

	Pathology of Nodules	**No. (%)**
Benign (*n* = 499)	Nodular hyperplasia	428 (85.8)
	Follicular adenoma	70 (14.0)
	Unidentified adenoma	1 (0.2)
Malignant (*n* = 324)	Papillary thyroid carcinoma	296 (91.4)
	Follicular thyroid carcinoma	15 (4.6)
	Medullary thyroid carcinoma	5 (1.5)
	Anaplastic carcinoma	5 (1.5)
	Others	3 (1.0)

* SD = standard deviation.

2.2. Testing CSF Detection Accuracy

The AmCAD-UT software device was developed to assist clinicians in analyzing the regions of interest (ROI) on the thyroid sonograms. Figure 1a shows the interface of the AmCAD-UT software. The software produced five computerized sonographic features (CSFs), namely, anechoic areas, hyperechoic foci, hypoechoic patterns, heterogeneous textures, and indistinct margins, providing their quantified values and visualizing them, using colors, to assist clinical differentiation of thyroid nodules. The quantification and visualization algorithms of AmCAD-UT have been disclosed in previous studies [20–24].

To validate the feature detection, we tested the quantified value of each sonographic feature against the ground truth, indicating the presence or absence of each feature on the sonogram (Table 2). The ground truth was determined by a panel of three thyroid specialists. All three specialists, certified to interpret the sonograms, with an average of 8.6 years (range 8–10 years) of experience and an average of 1366 readings (ranging from 700 to 2400 readings), independently read the sonograms to define the ROIs of the nodules and to determine the presence or absence of each feature. In cases of a discrepancy in interpretation, consensus was achieved by discussions among the panelists.

Table 2. Ground truth established by the panel of thyroid ultrasound specialists.

Sonographic Features	Determined by a Panel of Specialists	PH No. (%)	GE No. (%)	AL No. (%)
Anechoic Areas	Absence	68 (52.3%)	82 (82.0%)	96 (78.7%)
	Presence	62 (47.7%)	18 (18.0%)	26 (21.3%)
Hyperechoic Foci	Absence	97 (74.6%)	74 (74.0%)	93 (76.2%)
	Presence	33 (25.4%)	26 (26.0%)	29 (23.8%)
Hypoechoic Pattern	Absence	39 (30.0%)	30 (30.0%)	21 (17.2%)
	Presence	91 (70.0%)	70 (70.0%)	101 (82.8)
Heterogeneous Texture	Absence	11 (8.5%)	5 (5.0%)	10 (8.2%)
	Presence	119 (91.5%)	95 (95.0%)	112 (91.8%)
Indistinct Margin	Absence	102 (78.5%)	77 (77.0%)	49 (44.5%)
	Presence	28 (21.5%)	23 (23.0%)	61 (55.5%)

For all sonograms, the CAD device automatically performed detection and quantification of each feature and electronically stored the values in a database for subsequent analysis. The quantified values between nodules with the feature present and those without the feature were compared using a *t*-test and a *p*-value less than 0.05 was considered statistically significant. Since each feature was quantified as a continuous value to indicate a higher likelihood of presence by a larger value of the computerized feature, the receiver operating characteristic (ROC) curve for each feature was also generated againstthe corresponding "ground truth", with the area under the ROC curves (AUC) calculated to represent the detection accuracy. If the detection accuracy was 100%, i.e., AUC = 1.0, it meant a cut-off point can be found for the CSF to determine the presence of the sonographic feature, such that the findings by the computer and by the specialist panel were in 100% agreement. The higher the AUC value, the higher the agreement between the CSF and the ground truth. The statistical analysis was performed using MedCalc version 10.6.0.0 (MedCalc Software Ltd, Ostend, Belgium).

2.3. Assessing Diagnosis Performance of Readers Assisted with CSF

To assess whether the assistance of CSF provided by the CAD device can improve the readers' diagnosis accuracy, a multiple-reader multiple-case (MRMC) study [29–31] was performed. We recruited 18 clinicians (readers) to read 130 thyroid nodule sonograms. All the clinicians were licensed for an average of 8.78 years (ranging from 1 to 25 years experience) to perform ultrasound scans and interpret the sonograms, but had no experience in using AmCAD-UT. No reader had foreknowledge of the corresponding pathology results (benign or malignant) of the thyroid nodules. We first trained the readers with 20 training sonograms and the corresponding pathology results (10 benign or 10 malignant) to get the readers familiarized with the interface of the CAD software. Then, each of the 18 readers read each of 130 sonograms first without the CSF and then sequentially read the sonogram with the CSF provided [30]. The sequential reading with the CSF was to mimic how the CAD device would be used in clinical practice where the CSF information was provided as an integral part of the clinical reading and interpretation of sonograms. In other words, a reader read and scored every sonogram twice—one without CSF and one with CSF—for all 130 images. The order of the 130 sonograms was randomized and different for every reader. The scoring was scaled from 0 to 100 (0 = absolute benign, 100 = absolute malignant). We used the DBM MRMC 2.32 software (based on the Dorfman–Berbaum–Metz method) for generation of the ROC curves based on bi-normal models and for estimates of random effects of readers and cases [29,32–34]. Paired ROC curves of readers' scoring without and with assistance of CSF were generated against pathology, and the paired AUCs calculated. We also performed subgroup analysis for readers with different levels of experience. Readers with more than 6 years' experience of interpreting sonograms were referred to as senior readers and the others were called junior readers. We calculated the paired AUC for the two groups separately and compared the paired AUCs between the two groups.

3. Results

3.1. Detection Accuracy of Computerized Sonographic Features

The detection accuracy of each computerized feature was tested against the ground truth (Table 2). Regardless of the ultrasound scanner types, the difference between the two groups of quantified values (with or without presence of anechoic areas, hyperechoic foci, hypoechoic pattern, heterogeneous texture and indistinct margin) were significant, ranging from a *p*-value < 0.0001 to a *p*-value = 0.0347 (Table 3). The lowest agreement (*p*-value = 0.0045~0.0347) was observed in detecting the heterogeneous texture and the highest agreement between the computerized values and panel findings was observed for the anechoic area and the hyperechoic foci detection.

Table 3. Test results of the computerized sonographic features' detection accuracy.

		Anechoic Areas	**Hyperechoic Foci**	**Hypoechoic Pattern**	**Heterogeneous Texture**	**Indistinct Margin**
PH (*n* = 130)	AUC (95% CI)	0.902 (0.838–0.947)	0.913 (0.850–0.955)	0.837 (0.762–0.896)	0.701 (0.614–0.778)	0.702 (0.616–0.779)
	p-value	<0.0001	<0.0001	<0.0001	0.0347	0.0007
GE (*n* = 100)	AUC (95% CI)	0.888 (0.809–0.942)	0.825 (0.736–0.894)	0.847 (0.761–0.911)	0.627 (0.525–0.722)	0.766 (0.670–0.845)
	p-value	0.0001	<0.0001	<0.0001	0.0323	0.0002
AL (*n* = 122, *n* = 110 for Margin)	AUC (95% CI)	0.946 (0.890–0.979)	0.830 (0.751–0.892)	0.812 (0.732–0.877)	0.77 (0.685–0.841)	0.676 (0.580–0.762)
	p-value	<0.0001	0.0002	0.0156	0.0045	0.0002

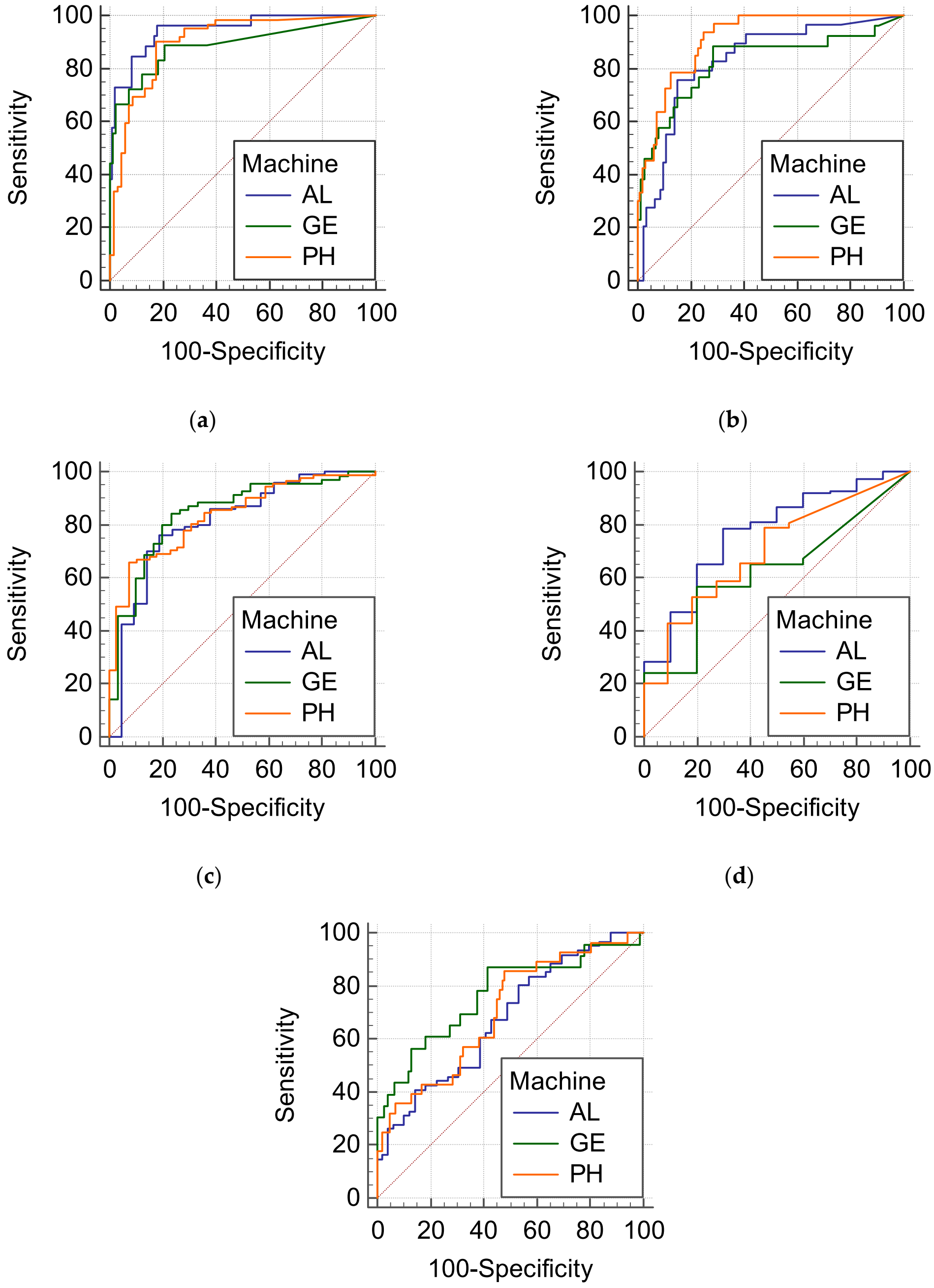
PH: Philips HDI 5000; GE: GE Voluson 730; AL: ALOKA Prosound2.

Similarly, in terms of the AUC against the ground truth, the detection accuracies of the quantified anechoic areas, hyperechoic foci, echogenicity (hypoechoic pattern), heterogeneous texture and indistinct margin were 0.946–0.888, 0.825–0.913, 0.812–0.847, 0.627–0.77 and 0.676–0.766 for various ultrasound scanners (Table 3 and Figure 3), respectively. The detection accuracies of the quantified heterogeneous texture and indistinct margin appeared to be lower than that of the quantified anechoic areas, hyperechoic foci, and hypoechoic pattern. These AUCs also indicated good agreement between the computerized feature values and the panel's readings.

3.2. Reader Performance Assisted with CSF

The reader performance using the computerized sonographic features generated by the CAD device was tested against the corresponding pathology. The accuracy of the readers in diagnosing malignant thyroid nodule sonograms was evaluated by the AUC based on the readers' scoring. AUCs without and with the computerized sonographic features (CSF) for each reader are shown in Table 4a. The mean AUC with CSF was significantly greater (*p*-value = 0.0420) than that without CSF using the DBM MRMC test, as shown in Table 4b. Moreover, the statistics showed that the difference was mainly observed for the junior readers (*p*-value = 0.1462 and 0.0265, respectively, for the senior and junior readers).

Figure 4 shows the graph of the paired ROC curves. Differences could be observed between the ROCs with and without CSF for all readers, indicating improvement in the reader's performance with the assistance of computerized features. In particular, Figure 4c showed the distinct ROC curves for the junior readers. However, only a slight difference was found in Figure 4b between the pair of ROC curves for senior readers. This demonstrated that the computerized features were especially beneficial to junior clinicians to supplement their relatively less experience in making diagnosis. Though an improvement was also observed for the senior readers assisted by the computerized features, a bigger sample size of senior readers might be required to show the significance.



(**a**)

(**b**)

(**c**)

(**d**)

(**e**)

Figure 3. ROC of detecting the presence by the computerized features: (**a**) anechoic area; (**b**) hyperechoic foci; (**c**) hypoechoic pattern; (**d**) heterogeneous texture; (**e**) indistinct margin.

Table 4. Reader performance in terms of AUC without and with CSF: (**a**) individual readers; (**b**) all readers, senior readers, and junior readers.

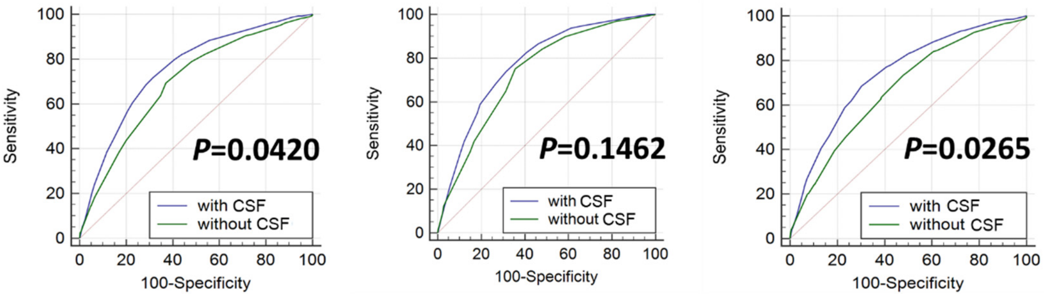
(a) AUC of Each Reader

		Seniority ** (Year)	AUC	
			Without CSF	**With CSF**
Senior * Readers	Reader 1	25	0.761	0.805
	Reader 2	21	0.786	0.821
	Reader 3	21	0.766	0.804
	Reader 8 #	12	0.807	0.810
	Reader 11 ##	7	0.832	0.825
	Reader 12 ##	7	0.822	0.817
	Reader 13	15	0.731	0.767
	Reader 14	10	0.614	0.738
	Reader 15	11	0.711	0.805
Junior Readers	Reader 4	5	0.740	0.770
	Reader 5	5	0.609	0.749
	Reader 6	5	0.744	0.753
	Reader 7	2	0.739	0.797
	Reader 9	1	0.514	0.642
	Reader 10	5	0.717	0.781
	Reader 16	3	0.653	0.761
	Reader 17	2	0.633	0.716
	Reader 18	1	0.784	0.810

(b) Mean AUC

	Without CSF (95% CI)	**With CSF (95% CI)**	**Improvement (95% CI)**	***p*-Value**
All Readers	0.720 (0.661, 0.780)	0.776 (0.708, 0.844)	0.056 (0.002, 0.110)	0.0420
Senior Readers	0.759 (0.706, 0.812)	0.799 (0.732, 0.866)	0.040 (−0.014, 0.094)	0.1462
Junior Readers	0.681 (0.608, 0.755)	0.753 (0.679, 0.827)	0.072 (0.009, 0.136)	**0.0265**

* Senior Readers: Seniority > 6 years; ** The item indicates the years of the reader been certified as a physician and able to use the ultrasound machine and interpret the sonograms; # The reader is a board-certified thyroid specialist; ## The reader is a board-certified radiologist; CI: confidence interval; Improvement = With CSF − Without CSF.



(**a**) (**b**) (**c**)

Figure 4. ROC curves with and without CSF: (**a**) all readers; (**b**) senior readers; (**c**) junior readers.

4. Discussion

4.1. Population and Features

The distribution and percentage of thyroid cancers were similar to that reported internationally [35–37], representing actual clinical practice and the demographic distribution of the studied population. Therefore, the possible effect of the nodule types and their distributions should be considered negligible. In addition, because the CSFs were found highly agreeable with the clinicians' findings for all three types of ultrasound scanners included in this study, the feature computerization of the CAD device should be generalizable to thyroid sonograms acquired for a general patient population in practical clinical situations.

4.2. CSF Accuracy

The sonographic features evaluated by the CAD device were tested to be significantly in agreement with the specialist's judgement (ground truth). The heterogeneous texture feature appeared to be less agreeable and was possibly due to the extremely unbalanced reading results by the specialist panel, where only 5–8.5% of nodules were not deemed to be heterogeneous. In clinical practice, even a slight unsmooth trace within a nodule would lead to a judgement of a heterogeneous nodule. In contrast, the CAD software provided the quantified values evaluating the degree of heterogeneity, which was more functional, with a continuous severity rating. Another less agreeable feature was the indistinct margin. An indistinct margin was supposedly indicated for a possible infiltration by a malignant nodule. However, determination of indistinct margins was highly subjective and greatly influenced by the clarity of the sonogram and the aim angle of the ultrasound probe. The readings of indistinct margins were thus with a relatively high variability and appeared to be less reliable. Since the CSFs were produced by the CAD software independently without interaction with clinicians, they could serve as second opinions to assist clinical diagnosis.

4.3. Reader Performance with CSF

An MRMC analysis was necessary to test the performance of a CAD device, since the radiological device was to assist the physicians and had to been shown of value to the physicians in making diagnosis. The MRMC study design offered a comprehensive analysis of the role of the device in actual clinical situations where clinicians of different experiences might use the device for reading a variety of cases [31]. Both the reader variability and interactions between readers and cases were considered by the DBM model in the MRMC analysis [34]. The scoring by rating the malignancy potentials of the nodules reflected the readers' interpretation confidence without and with the device's assistance. In this study, a total of 18 readers were recruited for the test and all readers received training by reading 20 sonograms prior to the test. The test sonograms were randomly selected from a sufficiently large database to avoid possible sampling bias. The analysis results demonstrated that the reading performances of clinicians, in terms of AUC, were significantly improved when assisted with the CSFs produced by the CAD. We had also observed that the greatest improvements were mainly made by the junior readers, although the higher diagnosis accuracy, with an average AUC near 0.8, was achieved by the senior readers. The CSFs appeared to be of value to both junior and senior readers, with the accuracy approaching and not exceeding 0.8. Since the improvement made by the senior readers was only about 5% (from AUC = 0.759 to 0.799), a larger number of senior readers might be needed to prove the statistical significance.

4.4. CAD Device in Clinical Practice

Major advantages of the CAD device included providing a critical second opinion, reducing time-consuming procedures, and avoiding oversight and interobserver variation [13]. A CAD device could serve as a reliable second opinion because it provides accurate information of the sonographic features in concordance with the ground truth. Independent double reading by another clinician was an alternative way to improve the detection accuracy [38]. However, independent double reading was a labor-demanding work for clinicians and might not be as efficient and effective as the assisted reading of a CAD device [19]. Furthermore, the sonographic reading of a human clinician was greatly dependent on one's past experience and often subject to momentary feel. In contrast, the CAD device could provide consistent and objective evaluation regardless of the time and environment settings, and thus could further reduce the interobserver variation [39].

Several CAD algorithms had been proposed for thyroid ultrasound images by other authors [25–28]. Acharya et al. characterized the thyroid nodules into benign and malignant classes using a combination of texture and discrete wavelet transform [25]. Chang et al. used the support vector machine classifier to differentiate malignant and benign nodules and showed accuracy similar to that obtained via visual inspection by radiologists [26]. Choi et al. suggested that the sensitivity of the junior reader was as good as that of the senior readers by using a CAD device with a lower the specificity [27]. Most recently, Wu et al. reported an MRMC study that showed a significant improvement in reading performance with assistance of the CAD device after a washout period [28]. Although the above studies assessed the differentiation of malignant and benign nodules by using the CAD device, this study would be the first study to evaluate the CAD device's performance in reading the sonographic features through comparison to readings by a panel of specialists and to assess the CSFs' effects on the reader performance via an MRMC study of sequential readings with and without CSF.

4.5. Study Limitation

This study had limitations regarding the sonographic features evaluated. First, hyperechoic foci were not further differentiated into microcalcifications, coarse calcifications, rim-shape calcifications, or colloid [20]. Second, the echogenicity level determined by the panel was not further classified into the four levels, namely, hyperechoic, isoechoic, mildly hypoechoic, and markedly hypoechoic [23], commonly used in clinical practice. Third, though the sequential readings without and with the CSF mimicked how the CAD device would be used in clinical practice but the sequential readings without a washout period might result in a recollection bias and impair the effect of the assistance by the CAD device.

In conclusion, the CAD device visualized and quantified the thyroid sonographic features in high concordance with a specialist panel and was shown to significantly improve clinicians' reading of the nodule's malignancy risk.

Author Contributions: Conceptualization, H.-C.T., K.-Y.C., M.-H.W., K.-J.C., C.-N.C. and A.C.; methodology, H.-C.T., K.-Y.C., M.-H.W. and A.C.; software, A.C.; validation, H.-C.T., K.-Y.C., M.-H.W. and A.C.; formal analysis, H.-C.T., K.-Y.C., M.-H.W. and A.C.; investigation, H.-C.T., K.-Y.C., M.-H.W. and A.C.; resources, K.-J.C., C.-N.C. and A.C.; data curation, K.-Y.C., M.-H.W., C.-N.C. and A.C.; writing—original draft preparation, H.-C.T. and A.C.; writing—review and editing, A.C.; visualization, A.C.; supervision, C.-N.C. and A.C.; project administration, C.-N.C.; funding acquisition, K.-J.C., C.-N.C. and A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Technology of Taiwan, grant number NSC 100-3114-B-002-004 and by the industry-university cooperative research project supported by AmCad BioMed Co.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of National Taiwan University Hospital (200805039R).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to thank Emma Shen for her help in running the AmCAD-UT software and performing the statistical analysis.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics. *CA Cancer J. Clin.* **2015**, *65*, 5–29. [CrossRef] [PubMed]
2. Chan, B.K.; Desser, T.S.; McDougall, I.R.; Weigel, R.J.; Jeffrey, R.B., Jr. Common and uncommon sonographic features of papillary thyroid carcinoma. *J. Ultrasound Med.* **2003**, *22*, 1083–1090. [CrossRef] [PubMed]

3. Papini, E.; Guglielmi, R.; Bianchini, A.; Crescenzi, A.; Taccogna, S.; Nardi, F.; Panunzi, C.; Rinaldi, R.; Toscano, V.; Pacella, C.M. Risk of malignancy in nonpalpable thyroid nodules: Predictive value of ultrasound and color-Doppler features. *Int. J. Clin. Endocrinol. Metab.* **2002**, *87*, 1941–1946. [CrossRef] [PubMed]
4. Haugen, B.R.; Alexander, E.K.; Bible, K.C.; Doherty, G.M.; Mandel, S.J.; Nikiforov, Y.E.; Pacini, F.; Randolph, G.W.; Sawka, A.M.; Schlumberger, M.; et al. 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* **2016**, *26*, 1–133. [CrossRef] [PubMed]
5. Tessler, F.N.; Middleton, W.D.; Grant, E.G.; Hoang, J.K.; Berland, L.L.; Teefey, S.A.; Cronan, J.J.; Beland, M.D.; Sesser, T.S.; Frates, M.C.; et al. ACR thyroid imaging, reporting and data system (TI-RADS): White paper of the ACR TI-RADS committee. *J. Am. Coll. Radiol.* **2017**, *14*, 587. [CrossRef]
6. Horvath, E.; Majlis, S.; Rossi, R.; Franco, C.; Niedmann, J.P.; Castro, A.; Dominguez, M. An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management. *Int. J. Clin. Endocrinol. Metab.* **2009**, *94*, 1748–1751. [CrossRef]
7. Russ, G. Risk stratification of thyroid nodules on ultrasonography with the French TI-RADS: Description and reflections. *Ultrasonography* **2016**, *35*, 25–38. [CrossRef]
8. Mechanick, J.I.; Harrell, R.M.; Allende-Vigo, M.Z.; Alvayero, C.; Arita-Melzer, O.; Aschner, P.; Camacho, P.M.; Castillo, R.Z.; Cerdas, S.; Coutinho, W.F.; et al. Transculturalization Recommendations for Developing Latin American Clinical Practice Algorithms in Endocrinology–Proceedings of the 2015 Pan-American Workshop by the American Association of Clinical Endocrinologists and American College of Endocrinology. *Endocr. Pract.* **2016**, *22*, 476–501. [CrossRef]
9. Seo, H.; Na, D.G.; Kim, J.H.; Kim, K.W.; Yoon, J.W. Ultrasound-Based Risk Stratification for Malignancy in Thyroid Nodules: A Four-Tier Categorization System. *Eur. Radiol.* **2015**, *25*, 2153–2162. [CrossRef]
10. Kwak, J.Y. Indications for fine needle aspiration in thyroid nodules. *Endocrinol. Metab.* **2013**, *28*, 81–85. [CrossRef]
11. Choi, S.H.; Kim, E.K.; Kwak, J.Y.; Kim, M.J.; Son, E.J. Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. *Thyroid* **2010**, *20*, 167–172. [CrossRef] [PubMed]
12. Park, C.S.; Kim, S.H.; Jung, S.L.; Kang, B.J.; Kim, J.Y.; Choi, J.J.; Sung, M.S.; Yim, H.W.; Jeong, S.H. Observer variability in the sonographic evaluation of thyroid nodules. *J. Clin. Ultrasound.* **2010**, *38*, 287–293. [CrossRef] [PubMed]
13. van Ginneken, B.; Schaefer-Prokop, C.M.; Prokop, M. Computer-aided diagnosis: How to move from the laboratory to the clinic. *Radiology* **2011**, *261*, 719–732. [CrossRef] [PubMed]
14. Petrick, N.; Sahiner, B.; Armato, S.G.; Bert, A.; Correale, L.; Delsanto, S.; Freedman, M.T.; Fryd, D.; Gur, D.; Hadjiiski, L.; et al. Evaluation of computer-aided detection and diagnosis systems. *Med. Phys.* **2013**, *40*, 087001. [CrossRef] [PubMed]
15. Giger, M.L.; Huo, Z.; Kupinski, M.A.; Vyborny, C.J. Computer-aided diagnosis in mammography. In *Handbook of Medical Imaging*; Fitzpatrick JMS, M., Ed.; SPIE PRESS: Bellingham, WA, USA, 2000; pp. 915–1004.
16. Singh, S.; Maxwell, J.; Baker, J.A.; Nicholas, J.L.; Lo, J.Y. Computer-aided classification of breast masses: Performance and interobserver variability of expert radiologists versus residents. *Radiology* **2011**, *258*, 73–80. [CrossRef]
17. Sugimoto, K.; Shiraishi, J.; Moriyasu, F.; Doi, K. Computer-aided diagnosis of focal liver lesions by use of physicians' subjective classification of echogenic patterns in baseline and contrast-enhanced ultrasonography. *Acad. Radiol.* **2009**, *16*, 401–411. [CrossRef]
18. Doi, K. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Comput. Med. Imaging Graph.* **2007**, *31*, 198–211. [CrossRef]
19. Jiang, Y.; Nishikawa, R.M.; Schmidt, R.A.; Metz, C.E. Comparison of independent double readings and computer-aided diagnosis (CAD) for the diagnosis of breast calcifications. *Acad. Radiol.* **2006**, *13*, 84–94. [CrossRef]
20. Chen, K.Y.; Chen, C.N.; Wu, M.H.; Ho, M.C.; Tai, H.C.; Huang, W.C.; Chung, Y.C.; Chen, A.; Chang, K.J. Computerized detection and quantification of microcalcifications in thyroid nodules. *Ultrasound Med. Biol.* **2011**, *37*, 870–878. [CrossRef]
21. Wu, M.H.; Chen, C.N.; Chen, K.Y.; Ho, M.C.; Tai, H.C.; Chung, Y.C.; Lo, C.P.; Chen, A.; Chang, K.J. Quantitative analysis of dynamic power Doppler sonograms for patients with thyroid nodules. *Ultrasound Med. Biol.* **2013**, *39*, 1543–1551. [CrossRef]
22. Chen, K.Y.; Chen, C.N.; Wu, M.H.; Ho, M.C.; Tai, H.C.; Kuo, W.H.; Huang, W.C.; Wang, Y.H.; Chen, A.; Chang, K.J. Computerized quantification of ultrasonic heterogeneity in thyroid nodules. *Ultrasound Med. Biol.* **2014**, *40*, 2581–2589. [CrossRef] [PubMed]
23. Wu, M.H.; Chen, C.N.; Chen, K.Y.; Ho, M.C.; Tai, H.C.; Wang, Y.H.; Chen, A.; Chang, K.-J. Quantitative analysis of echogenicity for patients with thyroid nodules. *Sci. Rep.* **2016**, *6*, 35632. [CrossRef] [PubMed]
24. Wu, M.H.; Chen, K.Y.; Chen, A.; Chen, C.N. Software-Based Analysis of the Taller-Than-Wide Feature of High-Risk Thyroid Nodules. *Ann. Surg. Oncol.* **2021**, *28*, 4347–4357. [CrossRef] [PubMed]
25. Acharya, U.R.; Faust, O.; Sree, S.V.; Molinari, F.; Suri, J.S. ThyroScreen system: High resolution ultrasound thyroid image characterization into benign and malignant classes using novel combination of texture and discrete wavelet transform. *Comput. Methods Programs Biomed.* **2012**, *107*, 233–241. [CrossRef] [PubMed]
26. Chang, Y.; Paul, A.K.; Kim, N.; Baek, J.H.; Choi, Y.J.; Ha, E.J.; Lee, K.D.; Lee, H.S.; Shin, D.; Kim, N. Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: A comparison with radiologist-based assessments. *Med. Phys.* **2016**, *43*, 554. [CrossRef]
27. Choi, Y.J.; Baek, J.H.; Park, H.S.; Shim, W.H.; Kim, T.Y.; Shong, Y.K.; Lee, J.H. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: Initial clinical assessment. *Thyroid* **2017**, *27*, 546–552. [CrossRef]

28. Wu, M.-H.; Chen, K.-Y.; Shih, S.-R.; Ho, M.C.; Tai, H.C.; Chang, K.J.; Chen, A.; Chen, C.N. Multi-reader multi-case study for performance evaluation of high-risk thyroid ultrasound with computer-aided detection. *Cancers* **2020**, *12*, 373. [CrossRef]
29. Hillis, S.L.; Berbaum, K.S. Power estimation for the Dorfman-Berbaum-Metz method. *Acad. Radiol.* **2004**, *11*, 1260–1273. [CrossRef]
30. Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in Premarket Notification (510(k)) Submissions. Available online: https://www.fda.gov/media/77642/download (accessed on 29 May 2022).
31. Gallas, B.D.; Chan, H.P.; D'Orsi, C.J.; Dodd, L.E.; Giger, M.L.; Gur, D.; Krupinski, E.A.; Metz, C.E.; Myers, K.J.; Obuchowski, N.A.; et al. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. *Acad. Radiol.* **2012**, *19*, 463–477. [CrossRef]
32. Metz ROC Software. Available online: http://metz-roc.uchicago.edu/MetzROC/software (accessed on 29 May 2022).
33. Xiao-Hua, Z.N.; Donna, K. McClish. In *Statistical Methods in Diagnostic Medicine*, 2nd ed.; John Wiley and Sons: Hoboken, NJ, USA, 2002.
34. Dendumrongsup, T.; Plumb, A.A.; Halligan, S.; Fanshawe, T.R.; Altman, D.G.; Mallett, S. Multi-reader multi-case studies using the area under the receiver operator feature curve as a measure of Diagnosis-Accuracy: Systematic review with a focus on quality of data reporting. *PLoS ONE* **2014**, *9*, e116018. [CrossRef]
35. Cancer Stat Facts: Thyroid Cancer. Available online: https://seer.cancer.gov/statfacts/html/thyro.html (accessed on 29 May 2022).
36. Moon, W.J.; Jung, S.L.; Lee, J.H.; Na, D.G.; Baek, J.H.; Lee, Y.H.; Kim, J.; Kim, H.S.; Byun, J.S.; Lee, D.H. Benign and malignant thyroid nodules: US differentiation–multicenter retrospective study. *Radiology* **2008**, *247*, 762–770. [CrossRef] [PubMed]
37. Taiwan Cancer Registryy. Available online: https://www.hpa.gov.tw/Pages/ashx/File.ashx?FilePath=~{}/File/Attach/6069/File_5962.pdf (accessed on 29 May 2022).
38. Thurfjell, E.L.; Lernevall, K.A.; Taube, A.A. Benefit of independent double reading in a population-based mammography screening program. *Radiology* **1994**, *191*, 241–244. [CrossRef] [PubMed]
39. Krupinski, E.A. Computer-aided detection in clinical environment: Benefits and challenges for radiologists. *Radiology* **2004**, *231*, 7–9. [CrossRef] [PubMed]