The Journal of Clinical Endocrinology & Metabolism, 2024, **109**, 2872–2881
https://doi.org/10.1210/clinem/dgae254
Advance access publication 13 April 2024
Clinical Research Article

Deep Learning Analysis With Gray Scale and Doppler Ultrasonography Images to Differentiate Graves' Disease

Han-Sang Baek,[1] Jinyoung Kim,[2] Chaiho Jeong,[1] Jeongmin Lee,[3] Jeonghoon Ha,[4] Kwanhoon Jo,[5] Min-Hee Kim,[3] Tae Seo Sohn,[1] Ihn Suk Lee,[6] Jong Min Lee,[6] and Dong-Jun Lim[4]

[1]Division of Endocrinology and Metabolism, Department of Internal Medicine, Uijeongbu St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Uijeongbu 11765, Republic of Korea
[2]Division of Endocrinology and Metabolism, Department of Internal Medicine, Yeouido St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul 07345, Republic of Korea
[3]Division of Endocrinology and Metabolism, Department of Internal Medicine, Eunpyeong St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul 03312, Republic of Korea
[4]Division of Endocrinology and Metabolism, Department of Internal Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea
[5]Division of Endocrinology and Metabolism, Department of Internal Medicine, Incheon St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Incheon 21431, Republic of Korea
[6]Division of Endocrinology and Metabolism, Department of Internal Medicine, Daejeon St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Daejeon 34943, Republic of Korea

Correspondence: Dong-Jun Lim, MD, PhD, Division of Endocrinology and Metabolism, Department of Internal Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 222, Banpo-daero, Seocho-gu, Seoul, 06591, Republic of Korea. Email: ldj6026@catholic.ac.kr.

Abstract

Context: Thyrotoxicosis requires accurate and expeditious differentiation between Graves' disease (GD) and thyroiditis to ensure effective treatment decisions.

Objective: This study aimed to develop a machine learning algorithm using ultrasonography and Doppler images to differentiate thyrotoxicosis subtypes, with a focus on GD.

Methods: This study included patients who initially presented with thyrotoxicosis and underwent thyroid ultrasonography at a single tertiary hospital. A total of 7719 ultrasonography images from 351 patients with GD and 2980 images from 136 patients with thyroiditis were used. Data augmentation techniques were applied to enhance the algorithm's performance. Two deep learning models, Xception and EfficientNetB0_2, were employed. Performance metrics such as accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score were calculated for both models. Image preprocessing, neural network model generation, and neural network training results verification were performed using DEEP:PHI® platform.

Results: The Xception model achieved 84.94% accuracy, 89.26% sensitivity, 73.17% specificity, 90.06% PPV, 71.43% NPV, and an F1 score of 89.66 for the diagnosis of GD. The EfficientNetB0_2 model exhibited 85.31% accuracy, 90.28% sensitivity, 71.78% specificity, 89.71% PPV, 73.05% NPV, and an F1 score of 89.99.

Conclusion: Machine learning models based on ultrasound and Doppler images showed promising results with high accuracy and sensitivity in differentiating GD from thyroiditis.

Key Words: thyrotoxicosis, Graves' disease, thyroiditis, ultrasonography, neural networks computer, artificial intelligence

Thyrotoxicosis presents a spectrum of clinical manifestations caused by an excessive increase in thyroid hormone levels in the body (1). Thyrotoxicosis can be caused by excessive secretion of thyroid hormones resulting from a hyperfunctioning thyroid gland or destructive thyroiditis resulting from the inflammation of the thyroid gland. Graves' disease (GD), an autoimmune disorder and one of the common causes of thyrotoxicosis, is caused by hyperfunctioning of the thyroid gland; meanwhile, subacute thyroiditis, painless thyroiditis, or autoimmune thyroiditis are forms of destructive thyroiditis (2).

The accurate and prompt differentiation of thyrotoxicosis subtypes is crucial for making appropriate treatment decisions (3). Although an antithyroid medication is necessary for treating GD, conservative treatment for symptom control is required for patients with destructive thyroiditis.

The differential diagnosis of thyroiditis and GD in clinical practice involves the measurement of serum TSH receptor antibody (TSH-R-Ab) levels and the performance of thyroid scintigraphy (99mTc pertechnetate or iodine-123) (3). Although thyroid scintigraphy is effective for diagnosing GD, it is only available in large hospitals or medical centers and is contraindicated during pregnancy (4, 5). TSH-R-Ab measurement provides an accurate diagnosis of GD with 98% to 99% sensitivity and specificity (3).

Received: 4 September 2023. **Editorial Decision:** 9 April 2024. **Corrected and Typeset:** 14 May 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of the Endocrine Society. All rights reserved. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from https://academic.oup.com/jcem/article/109/11/2872/7645064 by guest on 18 April 2025

Traditionally, thyroid ultrasonography was considered to be of limited clinical utility in patients with thyrotoxicosis. However, the importance of thyroid ultrasonography for immediate treatment decision-making has been recently emphasized owing to its noninvasive nature and real-time diagnostic capability (3, 6). Doppler ultrasonography can be used as an adjunctive method to assess thyroid gland characteristics and blood flow patterns for real-time diagnosis of GD. However, it has a modest sensitivity, at approximately 84%, and limited diagnostic utility (5, 7). Previous studies have highlighted a significant challenge in obtaining objective results due to high interobserver variability (8). Additionally, the current reliance on subjective interpretation in color Doppler ultrasonography presents difficulties in distinguishing between GD and autoimmune thyroiditis, as both conditions exhibit increased vascularity (2, 5). Therefore, it is necessary to improve the accuracy of ultrasonography and reduce interobserver variability.

In recent years, machine learning algorithms have emerged as promising tools for medical image analysis (9). An algorithm could be developed using machine learning to minimize interobserver variability and differentiate thyrotoxicosis based on blood flow patterns that are difficult for humans to distinguish (10). Previous machine learning-based studies using Doppler ultrasonography imaging reported a 10% improvement in diagnostic accuracy. However, these studies focused on detecting malignancies in thyroid nodules (11-13). Studies investigating the use of thyroid ultrasonography coupled with machine learning techniques for differentiating thyrotoxicosis are currently unavailable. Recent machine learning studies using thyroid scintigraphy images for differentiating thyrotoxicosis have reported diagnostic accuracies ranging from 73% to 97% (14).

Artificial intelligence (AI) has the potential to transform subjective diagnoses derived from expert physical examinations into more objective and quantifiable assessments. It can function as an intermediary tool, offering guidance and assistance during the interim period before a definitive diagnosis is established. Therefore, this study aimed to develop a machine learning algorithm capable of differentiating thyrotoxicosis subtypes with a specific focus on GD using Doppler ultrasonography images.

Materials and Methods

Study Population

In this retrospective study, we recruited patients with thyrotoxicosis who underwent thyroid ultrasonography between January 2010 and April 2022 at Seoul St. Mary's Hospital in Seoul, South Korea. Ultrasonography images stored in the Picture Archiving and Communication System (PACS) in Digital Imaging and Communications in Medicine format were used, and patient data (demographic characteristics and laboratory data) were collected from electronic medical records. The images were collected in a single tertiary hospital. Thyrotoxicosis was diagnosed using the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision codes E05–E06. These codes encompass various conditions such as thyrotoxicosis or hyperthyroidism, GD, acute thyroiditis, subacute thyroiditis, chronic thyroiditis with transient thyrotoxicosis, autoimmune thyroiditis, drug-induced thyroiditis, and other chronic thyroiditis. The medical records of 588 patients with suppressed TSH levels and free T4 values exceeding the upper limit of the reference range were reviewed. Following the exclusion of 43 patients diagnosed with thyroid nodule or cancer, 48 patients aged <19 years, 9 patients who underwent surgery or radioiodine ablation, and 1 patient diagnosed with hyperparathyroid adenoma, only 487 patients were included in the final analysis. The patients were classified into the GD and destructive thyroiditis groups based on clinical and laboratory findings. TSH-R-Ab served as the pivotal laboratory test for diagnosing GD; the specific disease was definitively confirmed by assessing the clinical course over a period of more than 3 months after ultrasound, as documented in the medical records. We did not use the ultrasonography findings to confirm the diagnosis as this may result in bias. Finally, 7719 thyroid ultrasonography images of 351 patients with GD and 2980 thyroid ultrasonography images of 136 patients with thyroiditis were used in this study. The ultrasonography images also included Doppler images. Of the 7719 thyroid ultrasonography images of patients with GD, 6937 (89.9%) were B-mode images, while 782 (10.2%) were Doppler images. Of the 2980 images of patients with thyroiditis, 2693 (90.3%) were B-mode images, while 287 (9.6%) were Doppler images. Approximately 22 images per patient, including right and left transverse and longitudinal views, were used for machine learning development. These ultrasonography images were obtained within 1 week before or after the date of thyrotoxicosis diagnosis. The image dataset was randomly divided in the order of train, validation, and test set in an 8:1:1 ratio (number of images of train: validation: test = 8561:1069:1069).

Data Protection and Privacy

Data used in this study were approved by the Catholic University Data Review Committee. All data were anonymized. All personal information was deleted when extracting images from PACS. The anonymized data were stored in encrypted files and computers to prevent patient reidentification. Due to the retrospective nature of this cohort study, there was no risk of physical or mental harm to the participants as a result of this study. Therefore, the requirement for obtaining informed consent was waived by the review board. This study adhered to the tenets of the Declaration of Helsinki and was approved by the Institutional Review Board of the Catholic University of Korea (KC22RASI0711).

Data Augmentation

To enhance the algorithm's performance and overcome limitations posed by the limited dataset size, data augmentation techniques were employed (15). Color images were converted to grayscale. The size of all images was standardized to 256 × 256. The range of pixel values across all images was unified using min-max normalization, preventing the neural network from learning unnecessary features. Gamma correction was used to improve image quality by adjusting the brightness, contrast, and color. To prevent the neural network from learning in areas other than the thyroid regions, corner detection, which identifies corners in the image, was used. Crop module was used to isolate the thyroid area. The same imaging preparing process was applied to all images in the dataset (Fig. 1).

Terminology, Model Training, and Experimental Environment

AI refers to the development of intelligent systems capable of undertaking tasks that typically require human intelligence (10).

Downloaded from https://academic.oup.com/jcem/article/109/11/2872/7645064 by guest on 18 April 2025

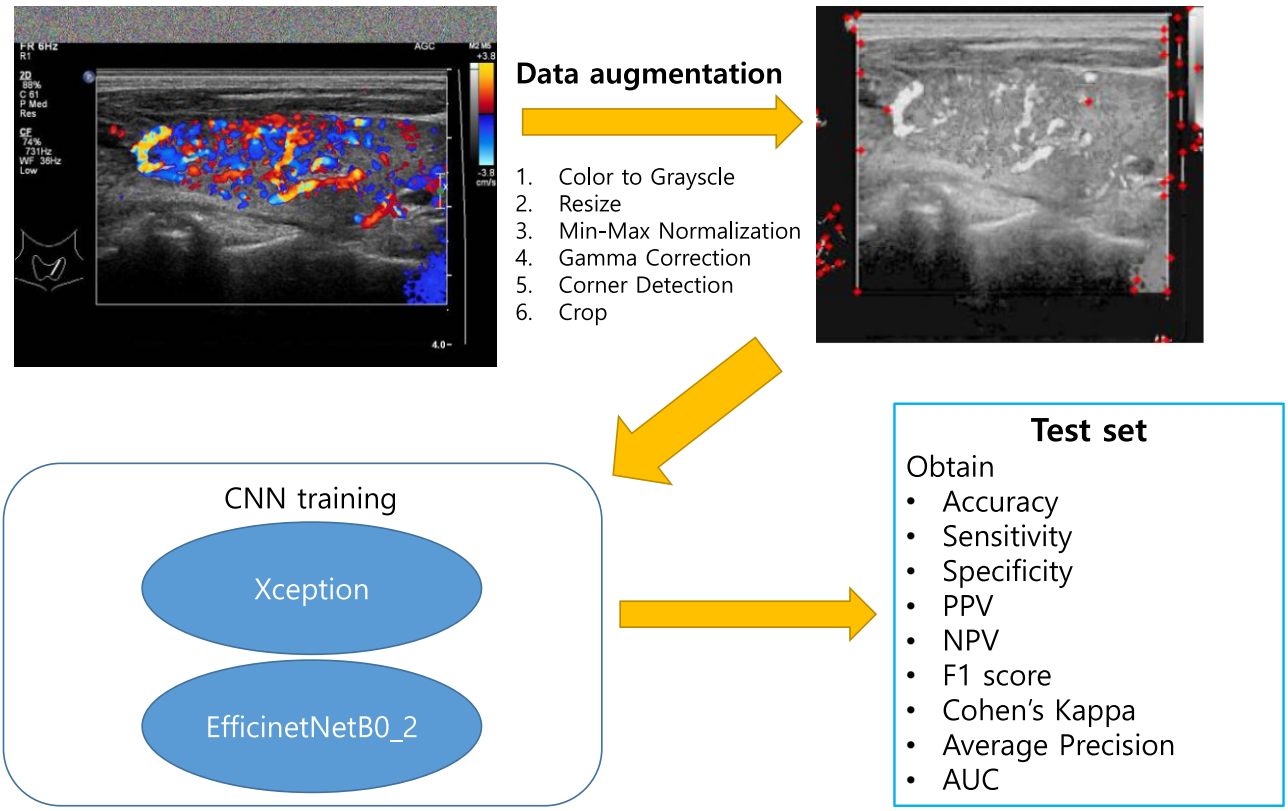

Figure 1. Image preprocessing. Data augmentation techniques were applied. Color images were converted to gray images. The size of all images was unified to 256 × 256. The range of pixel values of all images was then unified using min-max normalization to prevent the neural network from learning unnecessary features. Gamma correction was used to improve image quality through adjustment of brightness, contrast, and color. To prevent the neural network from learning areas other than the thyroid area, corner detection, which could find corners in the image, was used and the crop module was used to crop only the thyroid area. The same imaging preparing process was applied to all images in the dataset.

Machine learning is a subset of AI that focuses on the development of algorithms and models that enable computers to learn and make predictions or decisions without being explicitly programmed (16). Deep learning (DL) is a subfield of machine learning that utilizes neural networks with multiple layers to extract high-level features from raw data. DL models such as convolutional neural networks (CNNs) are particularly effective in image and speech recognition tasks as they can automatically learn hierarchical representations of data, leading to superior performance in various domains (9, 16).

This study aimed to confirm the usefulness of the early differentiation of thyrotoxicosis etiology using the DL program. For this purpose, 2 DL models, Xception and EfficientNetB0_2, were selected for our study considering their ability in image classification tasks. Both models were configured with a learning rate decay of 0.9 and a learning rate of 0.001, while the learning rate was set at 0.001. Subsequently, the diagnostic performance of each model in distinguishing GD or destructive thyroiditis was evaluated based on key metrics including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, Cohen's Kappa, average precision (AP), and area under the curve (AUC). Next, the models' ability to differentiate between GD and autoimmune thyroiditis or postpartum thyroiditis using the previously constructed two models was determined using only GD and autoimmune thyroiditis images, excluding subacute thyroiditis or other thyroiditis in dataset.

Introduced by Google's research team in 2016, Xception is a DL architecture derived from the Inception module. Similar to Inception V3 (DL model belonging to the Inception family of CNN) introduced in 2015, Xception focuses on enhancing the efficiency and accuracy of image classification tasks (17). By simultaneously learning channel mixing and spatial information, Xception strikes a balance between computational efficiency and high accuracy, making it suitable for various computer vision tasks, including image classification, object detection, and segmentation (18).

EfficientNet B0 to B7 developed by Google's research team in 2019 is a series of DL models designed to strike a balance between model size and performance in image classification tasks (19). Among these variants, EfficientNet B0_2 is a smaller variant with reduced depth, width, and resolution than B0, making it suitable for resource-limited scenarios such as edge devices or mobile applications. Despite its compact size, EfficientNet B0_2 maintains good performance and generalization capabilities.

In this study, DEEP:PHI was used to perform image preprocessing, neural network model generation, and neural network training results verification. DEEP:PHI platform is one of the platforms developed by DEEP-NOID, a Korean medical AI company comprising a graphical user interface (20). DEEP:PHI is a platform for AI model planning and solution development based on assembling modules (https://www.deepnoid.com/ai-platform?lang=en). Several DL programs can be learned in the DEEP:PHI environment, an AI platform, with Xception and EfficeintNET B0_2 demonstrating superior performance. No employees from DEEP:PHI were involved in the study design, data analysis, or manuscript preparation. Our

Downloaded from https://academic.oup.com/jcem/article/109/11/2872/7645064 by guest on 18 April 2025

team consists entirely of independent researchers with no financial or professional connections to DEEP:PHI, guaranteeing unbiased study findings driven solely by scientific inquiry.

Results

Demographic and Clinical Characteristics of the Study Population

The study population consisted of patients with GD (n= 351) and those with destructive thyroiditis (n = 136). Patients with destructive thyroiditis included those with autoimmune thyroiditis (n = 46), subacute thyroiditis (n = 31), amiodarone induced thyroiditis (n = 2), postpartum thyroiditis (n = 6), and other thyroiditis (n = 51). Patients with GD were slightly younger, and sex distribution was similar between the groups. Patients with GD had notably higher levels of free T4 and T3 but lower levels of TSH compared with patients with thyroiditis. No significant differences were observed in the thyroid peroxidase antibody or thyroglobulin antibody levels between the groups. Patients with GD had significantly higher thyrotropin receptor antibody (thyrotropin-binding inhibitor immunoglobulin) levels than those with thyroiditis (17.7 ± 19.5 IU/L vs 2.0 ± 5.6 IU/L, $P < .001$). There were also differences between the 2 groups in the range of thyrotropin-binding inhibitor immunoglobulin (10.65 [4.71-25.00] vs 0.36 [0.30-0.79], $P < .001$). Detailed information is presented in Table 1.

Performances of Machine Learning Models

The performance metrics of machine learning models Xception and EfficientNetB0_2 for differentiating thyrotoxicosis using Doppler ultrasonography images are presented in Table 2. The Xception model achieved an accuracy of 90.38% during the training phase, with a sensitivity of 92.37%, a specificity of 84.93%, a PPV of 94.39%, a negative NPV of 80.20%, and an F1 score of 93.37. During validation, the model achieved an accuracy of 82.86%, a sensitivity of 87.87%, a specificity of 69.03%, a PPV of 88.68%, an NPV of 67.32%, and an F1 score of 88.28. During testing, the Xception model demonstrated an accuracy of 84.94%, a sensitivity of 89.26%, a specificity of 73.17%, a PPV of 90.06%, an NPV of 71.43%, an F1 score of 89.66, a Cohen's Kappa coefficient of 61.95, an AP of 96.22, and an AUC of 90.48.

Table 1. Demographic and clinical characteristics of the study population

	Graves' (n = 351)	Thyroiditis (n = 136)	*P*
Age	43.3 ± 14.6	47.3 ± 13.6	.006
Female sex (%)	260 (74.1)	106 (77.9)	.442
Free T4 (0.89-1.79 ng/dL)	2.8 ± 1.1	2.4 ± 0.8	<.001
T3 (0.6-1.81 ng/dL)	2.6 ± 1.8	1.8 ± 0.8	<.001
TSH (0.55-4.05 uIU/mL)	0.1 ± 0.3	0.3 ± 0.8	.002
TPO-Ab (<60 IU/mL)	1871.9 ± 3415.8	1343.8 ± 3049.1	.251
Tg-Ab (<60 IU/mL)	475.4 ± 1373.3	296.4 ± 525.9	.227
TBII (<1.75 IU/L)	17.7 ± 19.5	2.0 ± 5.6	<.001

Abbreviations: TBII, thyrotropin-binding inhibitor immunoglobulin; Tg-Ab, thyroglobulin antibody; TPO-Ab, thyroid peroxidase antibody.

Similarly, the EfficientNetB0_2 model achieved an accuracy of 87.53% during training, with a sensitivity of 90.19%, a specificity of 79.65%, a PPV of 92.93%, an NPV of 73.25%, and an F1 score of 91.54. During validation, it achieved an accuracy of 83.51%, a sensitivity of 88.03%, a specificity of 71.32%, a PPV of 89.22%, an NPV of 68.86%, and an F1 score of 88.62. During testing, the EfficientNetB0_2 model exhibited an accuracy of 85.31%, a sensitivity of 90.28%, a specificity of 71.78%, a PPV of 89.71%, an NPV of 73.05%, an F1 score of 89.99, a Cohen's Kappa coefficient of 62.40, an AP of 96.63, and an AUC of 91.30.

These 2 models also showed high diagnostic accuracy in distinguishing GD from autoimmune thyroiditis or postpartum thyroiditis, excluding subacute thyroiditis or other types of thyroiditis. During testing, the Xception model showed an accuracy of 92.19%, a sensitivity of 94.33%, a specificity of 74.06%, a PPV of 96.86%, and a NPV of 60.62%. The EfficientNetB0_2 model showed an accuracy of 93.13%, a sensitivity of 95.13%, a specificity of 76.33%, a PPV of 97.10%, and a NPV of 65.0%. Further details are included in Table 3.

These performance metrics demonstrated the effectiveness of the Xception and EfficientNetB0_2 models in accurately differentiating thyrotoxicosis subtypes using Doppler ultrasonography images. The 2 models' high accuracies and sensitivities highlight their potential as valuable tools in clinical practice.

Discussion

The current study aimed to develop a machine learning algorithm capable of differentiating thyrotoxicosis subtypes with a specific focus on GD using ultrasonography images including Doppler images. Both models achieved high accuracy, sensitivity, and specificity during training, validation, and testing phases. The Xception model exhibited an accuracy of 84.94% during testing, while the EfficientNetB0_2 model achieved an accuracy of 85.31%. These results indicate that these machine learning models have the potential to accurately differentiate GD from thyroiditis based on Doppler ultrasonography images. Moreover, the F1 scores, Cohen's Kappa coefficients, AP, and AUC further support the robustness and overall performance of these machine learning models. These metrics demonstrate the ability of these models to balance precision and recall while maintaining their overall discriminative power.

The accuracy observed in this study was similar to or noninferior to that reported in previous studies using Doppler images to differentiate GD from other types of thyrotoxicosis (2, 21-23). In a previous study conducted by our research team using microvascular images, which are more advanced than conventional Doppler images, the accuracy in discriminating GD from other thyrotoxicosis was 85.2% (2).

However, the diagnosis of ultrasonography images based on machine learning offers additional benefits beyond just accuracy. The interpretation of ultrasonography images including Doppler can be subjective and may vary among different observers. AI could help minimize inter- and intraobserver variations and obtain standard images (10). This feature of AI is particularly valuable for assessors with limited experience in performing thyroid ultrasonography. Narange et al conducted a study where nurses with no prior ultrasonography experience followed AI guidance to scan patients with

Downloaded from https://academic.oup.com/jcem/article/109/11/2872/7645064 by guest on 18 April 2025

Table 2. Performance metrics of deep learning models

Model	Class	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 Score	Cohen's Kappa	AP	AUC
Xception	Train	90.38	92.37	84.93	94.39	80.20	93.37			
	Validation	82.86	87.87	69.03	88.68	67.32	88.28			
	Test	84.94	89.26	73.17	90.06	71.43	89.66	61.95	96.22	90.48
EfficientNetB0_2	Train	87.53	90.19	79.65	92.93	73.25	91.54			
	Validation	83.51	88.03	71.32	89.22	68.86	88.62			
	Test	85.31	90.28	71.78	89.71	73.05	89.99	62.40	96.63	91.30

Abbreviations: AP, average precision; AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

Table 3. Performance metrics of deep learning models for diagnosing Graves' disease from autoimmune thyroiditis or postpartum thyroiditis (excluding subacute thyroiditis or other thyroiditis in dataset)

Model	Class	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 Score	Cohen's Kappa	AP	AUC
Xception	Train	92.19	94.33	74.06	96.86	60.62	95.58			
	Validation	86.66	95.91	45.94	88.66	71.80	92.14			
	Test	89.23	98.16	33.06	90.22	74.07	94.02	40.69	97.96	90.97
EfficientNetB0_2	Train	93.13	95.13	76.33	97.12	65.08	96.12			
	Validation	88.53	93.50	49.69	93.55	49.50	93.53			
	Test	92.74	98.03	59.50	93.84	82.76	95.89	65.24	77.14	92.12

Abbreviations: AP, average precision; AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

a 10-view echocardiographic protocol and assessed the left ventricular size and function, right ventricular size, and presence of a pericardial effusion. The results were compared with those of expert echocardiographers who blindly reviewed the scans. The diagnostic accuracy was 92.5% for right ventricular size to 98.8% for pericardial effusion (24). Furthermore, a multicenter diagnostic study from China utilized a DL-based artificial intelligence model to assist in thyroid nodule diagnosis. The study reported that the diagnostic accuracy rates were 2.9% higher than that of a senior radiologist and 4.6% higher than that of a junior radiologist (11).

Although the number of machine learning studies using thyroid ultrasonography with a reported diagnostic accuracy of 64% to 96.1% has increased, all studies have predominantly focused on diagnosing thyroid cancer in thyroid nodules (12). Qiao et al conducted a study to distinguish between GD and subacute thyroiditis by constructing a DL model using thyroid scintigraphy images. They reported a diagnostic accuracy for GD of 88.6% to 92.8% (14). AI has primarily been applied to static images such as X-ray and computed tomography images, as applying AI to ultrasonography images with a dynamic nature poses considerable challenges. However, recent advancements, such as a CNN-based view classification of echocardiography, have expanded the applicability of AI to echocardiography (25-27). Accordingly, the use of AI in thyroid ultrasonography is expected to increase. Zhang et al developed a DL model to diagnose Hashimoto's thyroiditis based on thyroid ultrasonography video data, reporting an accuracy of approximately 90% (28).

One challenge encountered in conducting machine learning studies using ultrasonography images is how to effectively train Doppler color images. GD can be diagnosed using B-mode images, as approximately 70% of ultrasonography images of patients with GD have specific features such as diffuse hypoechogenicity and goiter (29); however, combined with Doppler images, the diagnostic rate increases by approximately 7% (23, 30). Accordingly, many studies included Doppler images when training a DL model (31). However, machine learning usually involves training using 1 channel, unifying gray images and color images (32), including the AI platform that is used in our study. Therefore, several previous studies used quantified values or color image information extracted from a specific region of interest when training Doppler images for DL. Moustafa et al extracted the vascularity index in manually drawn region of interests, using the color bar gradient for developing a machine learning model to detect breast cancer. They reported that the accuracy of the machine learning model increased between 92.5% and 95.8% when using color Doppler images (33). Zhu et al investigated the use of machine learning in the diagnosis of thyroid cancer based on thyroid nodules and found that using the "number of vascularity" increased the accuracy to approximately 92.5% (13). Furthermore, recently, Gomez-Flores et al trained a DL model by converting gray images into color images in the classification of breast lesions (34). Our DL model showed an accuracy of approximately 85% with scope for further improvement. A recent review suggested that a DL model trained using "multimodal images" may improve the diagnostic accuracy of DL models and overcome the limitations of using 1-channel analysis (32). Notably, a study on DL algorithms, developed using multimodal ultrasonography images, reported that malignant ovarian cancer could be distinguished from benign tumors using grayscale images and color Doppler images simultaneously with an accuracy of approximately 90% (35). When training DL algorithms to differentiate thyrotoxicosis, gray images and Doppler images (or the quantified vascularity scores) could be learned through multiple channels in the future (2, 31, 32).

Downloaded from https://academic.oup.com/jcem/article/109/11/2872/7645064 by guest on 18 April 2025

In addition to the methodological limitations when training DL, Doppler ultrasonography also has its own limitations. Currently, the main limitation in using Doppler ultrasonography lies in the differential diagnosis of thyrotoxicosis, specifically in differentiating GD from autoimmune thyroiditis. GD and autoimmune thyroiditis are typically characterized by increased thyroid vascularity with interferon-γ-inducible chemokines (36), resulting in a "false-positive" result where autoimmune thyroiditis may be misidentified as GD (2, 23). Despite its high sensitivity, the DL method developed in our study exhibited low specificity for differentiating GD and autoimmune disease.

However, our study focused on differentiating thyrotoxicosis, which precedes the differentiation between GD and autoimmune thyroiditis. The results of our study suggest that ultrasonography using DL can provide reliable clinical information quickly, particularly benefitting physicians with limited experience in managing patients with thyroid disease. In other words, although AI-based ultrasound cannot perfectly differentiate thyrotoxicosis due to the aforementioned limitations, it can serve as a tool to reassure patients in real time and assist doctors in determining patient management (5). With this approach, the likelihood of diagnosing GD could be quantified using AI-assisted ultrasound.

The high sensitivity of the DL models we developed suggests that these models can effectively identify patients with GD, minimizing the risk of obtaining false-negative results. Furthermore, the PPV values reflect the ability of these models to accurately predict the presence of GD. This high sensitivity can prove valuable in situations requiring immediate diagnosis of thyrotoxicosis, such as in the emergency department or primary care settings. In patients presenting with thyrotoxicosis, diagnosing based solely on free T4 and TSH levels presents some limitations in terms of establishing the etiology of thyrotoxicosis due to the time delays associated with additional TSH-R-Ab testing, facility constraints related to thyroid scintigraphy, and the limited diagnostic ability of thyroid ultrasonography. TSH-R-Ab testing has good diagnostic ability with 98% sensitivity and 99% specificity; however, testing may be time-consuming depending on the clinical setting, making real-time diagnosis difficult (37). AI-based ultrasonography can overcome these limitations, allowing the rapid diagnosis of thyrotoxicosis with diagnostic accuracy comparable to that reported for experienced doctors, facilitating prompt treatment decisions (medication therapy vs clinical observation). Imagine several scenarios where a less experienced doctor or a doctor with low medical facilities in an underdeveloped country is on night duty in the emergency room and must decide on the admission or discharge of a patient with thyrotoxicosis. In such situations, the laboratory may not have the capability to conduct TSH-R-Ab tests. In such situations, the doctor is compelled to depend on clinical judgment, symptoms, and the limited array of nonspecific tests at their disposal to promptly make a decision. They must recognize the importance of careful monitoring and may opt for a cautious approach by considering admission for further evaluation when more comprehensive testing is available. Automating the differentiation of thyrotoxicosis subtypes through these models can facilitate prompt treatment decisions, thus reducing the risk of inappropriate medication administration and the occurrence of associated complications. In this context, AI-assisted ultrasonography serves as a point-of-care testing tool for physicians lacking expertise in thyroid ultrasound interpretation, especially in environments like the emergency department.

Critics of ultrasonography use to differentiate thyrotoxicosis before confirming the TSH-R-AB results argue that ultrasonography may have minimal impact on patient treatment decisions; in situations where TSH-R-AB is not available, accessing an AI system for support may pose challenges. However, for physicians experienced in managing patients with thyrotoxicosis, the use of ultrasonography may have a negligible impact on treatment decisions. However, for less experienced practitioners, this tool could validate clinical assessments and potentially improve patient satisfaction. AI-driven diagnostic tools have the potential for real-time clinical applications beyond point-of-care settings. A study in Germany highlighted the utility of this approach with 93% of general physicians citing immediate decision-making and enhanced diagnostic certainty as key benefits of employing such tools (38). A pilot study from the United Kingdom reported that real-time ultrasound for thyroid nodules not only boosts physician confidence but also enhances patient satisfaction during examinations (39). A DL model trained by scintigraphy image provided diagnostic assistance to nuclear medicine residents and faster results; furthermore, the accuracy of the first-year resident in diagnosing GD improved from 87.3% to 92.1% with AI assistance (14). Furthermore, a DL model to diagnose Hashimoto's thyroiditis showed a higher accuracy than that achieved by radiologists (83.2% vs 79.8%) (28).

AI systems may be also beneficial in specific situations such as in the emergency department, particularly in challenging cases such as the diagnosis of severe thyrotoxic symptoms or thyrotoxic periodic paralysis (40). The integration of ultrasonography with AI systems is expected to expedite the diagnosis of GD, potentially reducing emergency department stays (41). Our cohort study, validated for thyrotoxicosis with thyroid function tests, suggests that our model aligns with real-world practices.

We assumed the scenario of managing GD for specificity. In standard outpatient management, thyroid function tests are initially conducted, followed by the administration of beta-blockers after detection of thyrotoxicosis, which is subsequently confirmed with antithyroid drug (ATD) following TSH-R-Ab positivity. Conversely, AI-supported management incorporates same-day real-time ultrasound with AI immediately after thyrotoxicosis detection, expediting the initiation of ATD and confirming TSH-R-Ab positivity thereafter. Similarly, in the emergency department, AI-supported management includes early initiation of ATDs facilitated by same-day real-time ultrasound with AI, potentially reducing emergency department stays (Fig. 2).

In addition, real-time testing using an AI system can be used to overcome challenges such as the unavailability of hospitals, lack of trained and skilled staff performing tests, and poor compliance with quality assurance protocols, particularly in rural or remote areas in low-income or developing countries (42, 43). To this end, further research should be conducted to examine the clinical utility and the effect on physician and patient satisfaction of AI-assisted point-of-care diagnosis.

This study has some limitations. First, it was a retrospective study conducted at a single center, which may introduce biases and limit the generalizability of study results. There may be differences in the ultrasonography machines used and the imaging protocols applied by different physicians or

Downloaded from https://academic.oup.com/jcem/article/109/11/2872/7645064 by guest on 18 April 2025

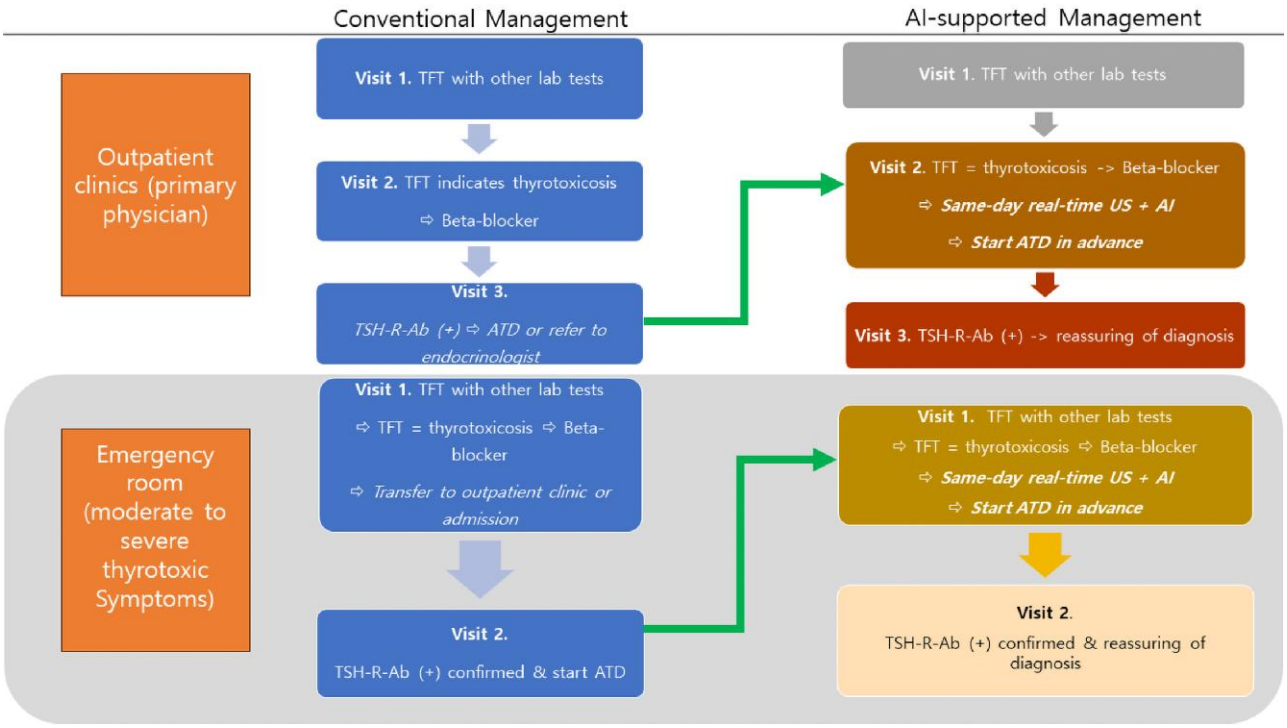

Figure 2. Hypothetical scenario of AI-supported US system in GD management. TFTs are performed initially, followed by the administration of beta-blockers after detection of thyrotoxicosis, which is subsequently confirmed with ATD following TSH-R-Ab positivity. Conversely, AI-supported management incorporates same-day real-time US with AI after thyrotoxicosis detection, expediting the initiation of ATD and providing subsequent confirmation of TSH-R-Ab positivity. Similarly, in the emergency department, AI-supported management includes early ATD initiation facilitated by same-day real-time US with AI, potentially reducing emergency department stays.

Abbreviations: AI, artificial intelligence; ATD, antithyroid drug; GD, Graves' disease; TFT, thyroid function tests; TSH-R-Ab, TSH receptor antibody; US, ultrasonography.

sonographers. As this was a retrospective study, patient personal information was deleted from the images downloaded from PACS, making it difficult to identify the ultrasound machine or probe used to obtain the images. Additionally, given that the training, validation, and test sets are divided within the same cohort according to a specific ratio, it remains unknown whether the currently trained program will show similar performance in other cohorts. Further validation with larger and more diverse datasets from multiple centers would be beneficial to determine the performance of these models in different populations. In addition, we recognize the imbalance in our study's dataset, with a 2:1 ratio of GD to thyroiditis images. Such an imbalance can impact model training. However, this imbalance was mitigated using data augmentation techniques, as suggested in previous literature (44). Dividing the dataset randomly without patient-based separation could lead to overfitting. Nonetheless, we have minimized this issue by applying various data augmentation techniques. Some studies reported that data augmentation can significantly alleviate overfitting (45), while other research indicated that data augmentation methods can reduce the extent of overfitting (46). Hence, future research should consider these findings and explore methods such as cross-validation to enhance the robustness of our model evaluation (47). Although the current phase of our research did not include external validation, future iterations of this work will seek external validation to ensure generalizability and reliability. This future study is aimed at assessing the model's performance across diverse patient populations and clinical settings, ensuring its robustness and applicability in real-world scenarios. In addition, the process of data splitting based on images in AI diagnostics could lead to issues like data leakage, overfitting, limited generalization, and eventually overestimation of AI's capability (48, 49). These challenges emphasize the need for meticulous data handling and the validation of models on external datasets to ensure their robustness and applicability in clinical settings (50). Future directions should focus on establishing clear guidelines for the safe implementation and assessment of AI technology, alongside robust empirical research to validate its benefits and understand its capabilities and limitations in real-world settings (51). Additionally, although these machine learning models aimed to minimize interobserver variability, further efforts to standardize image acquisition and interpretation protocols are needed to ensure consistent and reliable results. Moreover, more fine data augmentation processes may be necessary in future studies to enable DL models to detect the thyroid parenchyma more correctly. To overcome this issue, capturing moving images could be used in future studies (28). Lastly, using the open AI platform, finding accurate evidence for diagnostic elements or evidence for false-negative or false-positive results can be challenging, especially in the context of using AI in medical diagnostics, DEEP-NOID. In addition, although CNNs, including Xception and EfficientNet, are designed to utilize 3-channel (RGB) data to leverage the full spectrum of information in color images, we conducted a single-channel analysis. Due to this platform's limitations, our analysis necessitated the conversion of RGB images into grayscale images, effectively transforming them into single-channel data. This conversion was a prerequisite for analysis within the DEEP-NOID environment, which, as

Downloaded from https://academic.oup.com/jcem/article/109/11/2872/7645064 by guest on 18 April 2025

of our study's execution, supported only 1-channel input. Meanwhile, analyzing ultrasound images using a single-channel approach could offer some benefits over multichannel analysis. Single-channel ultrasound images are simpler and more computationally efficient to analyze, particularly beneficial for real-time analysis or managing large datasets (52). This simplicity can lead to reduced manufacturing costs and increased portability, which is particularly beneficial in resource-limited settings or for applications requiring compact and mobile solutions (53). By reducing the complexity of the data, single-channel imaging allows for a more focused examination of specific features or structures, eliminating the potential distractions from additional information present in multichannel images (52, 53). However, our approach might have seemed to oversimplify the complexity and versatility of CNNs in handling multichannel data (32). Moving forward, we aim to explore additional platforms and methodologies that can accommodate the intrinsic advantages of multichannel data analysis, further enhancing the accuracy and applicability of our findings in the field. Despite these limitations, to the best of our knowledge, this study is the first to identify thyrotoxicosis using AI-based ultrasonography. Future research should address dataset imbalances and refine our methodological approaches to ensure the robustness and generalizability of our AI models. We aim to enhance the applicability and reliability of AI in real-world clinical settings by prioritizing external validation and exploring more sophisticated data analysis methodologies.

We acknowledge the challenges and limitations of using AI in clinical practice, with concerns that its potential to contribute to misdiagnoses and inappropriate treatment decisions. Misdiagnosis can lead to inappropriate treatment, such as using methimazole in destructive thyroiditis, affecting patient outcomes negatively (54). However, as suggested in a previous review, rather than expecting AI to deliver flawless results, it should be viewed as a tool with the potential to improve outcomes continuously, particularly when integrated with physician expertise (51). AI has the potential to make subjective diagnoses, based on expert physical examinations, more objective and quantifiable; in addition, it can serve as a transitional method while awaiting confirmative diagnoses (51, 55-57). Thus, we advocate positioning AI as a supplementary tool aimed at bolstering clinicians' capabilities by providing data-driven insights to enhance treatment decisions, rather than supplanting their expertise (56). It is essential to emphasize the indispensable role of comprehensive clinical evaluations including laboratory values, clinical history, and physical examination in distinguishing between conditions like GD and thyroiditis (58, 59). AI significantly enhances diagnostic accuracy and patient care by providing insights and facilitating data interpretation, yet it operates within a diagnostic framework underpinned by healthcare professionals' expertise (51, 55, 56).

The practical implementation of AI in healthcare settings faces significant hurdles, including integration into existing workflows and the lack of evidence from clinical trials to prove its efficacy (51). For example, concerns may arise regarding the availability of ultrasonography in the emergency room. However, considering that the availability of ultrasound has improved in emergency rooms (60), thyroid ultrasound has already been used for evaluating patients with thyroid diseases (2, 61), and AI technology with thyroid ultrasound has been also employed (62). AI solutions could aid medical practice without disrupting existing process. For this purpose, according to Aung et al, it is important to adequately reflect the opinions of stakeholders and sufficiently educate physicians and healthcare providers on the use AI technology (51). Additionally, methods such as active learning and federated learning are suggested to improve the generalization of AI models (56). Despite these challenges, our manuscript advocates for the balanced integration of AI, emphasizing its role in augmenting rather than replacing the clinician's judgment. Future directions should focus on establishing clear guidelines for the safe implementation and assessment of AI technology, alongside robust empirical research to validate its benefits and understand its capabilities and limitations in real-world settings.

By underscoring AI's complementary role in healthcare, this study advocates for a thoughtful and effective integration of technology to support clinical practices, with the aim of improving patient care through enhanced diagnostic and treatment processes.

In conclusion, our study demonstrated the effectiveness of machine learning algorithms, specifically Xception and EfficientNetB0_2 models, in differentiating thyrotoxicosis subtypes using ultrasonography images. These models showed high accuracy and sensitivity, indicating their potential as valuable tools in clinical practice. Future studies should focus on prospective validation and integration of these models into clinical workflows to assess their real-world impact and usefulness in improving patient care.

Acknowledgments

We would like to thank the Catholic Information Convergence Institute and the data science team for the useful discussion of the experiment. A portion of this study was presented in abstract form at the Asia Oceanian Congress of Endocrinology and Seoul International Congress of Endocrinology and Metabolism 2023 in Seoul, Korea and the ENDO 2024, Boston, MA, USA.

Funding

There is no funding resource for this study.

Disclosures

The authors declare that they have no competing interest.

Data Availability

Some or all datasets generated during and/or analyzed during the current study are not publicly available but are available from the corresponding author upon reasonable request.

References

1. Badiu C. Williams textbook of endocrinology. *Acta Endocrinol (Buchar)*. 2019;15(3):416.
2. Baek HS, Park JY, Jeong CH, Ha J, Kang MI, Lim DJ. Usefulness of real-time quantitative microvascular ultrasonography for differentiation of Graves' disease from destructive thyroiditis in thyrotoxic patients. *Endocrinol Metab (Seoul)*. 2022;37(2):323-332.
3. Kahaly GJ. Management of Graves thyroidal and extrathyroidal disease: an update. *J Clin Endocrinol Metab*. 2020;105(12):3704-3720.

Downloaded from https://academic.oup.com/jcem/article/109/11/2872/7645064 by guest on 18 April 2025

4. Scappaticcio L, Trimboli P, Keller F, Imperiali M, Piccardo A, Giovanella L. Diagnostic testing for Graves' or non-Graves' hyperthyroidism: a comparison of two thyrotropin receptor antibody immunoassays with thyroid scintigraphy and ultrasonography. *Clin Endocrinol (Oxf)*. 2020;92(2):169-178.
5. Rosario PW, Santos JB, Nunes NS, da Silva AL, Calsolari MR. Color flow Doppler sonography for the etiologic diagnosis of thyrotoxicosis. *Horm Metab Res*. 2014;46(7):505-509.
6. Alzahrani AS, Ceresini G, Aldasouqi SA. Role of ultrasonography in the differential diagnosis of thyrotoxicosis: a noninvasive, cost-effective, and widely available but underutilized diagnostic tool. *Endocr Pract*. 2012;18(4):567-578.
7. Donkol RH, Nada AM, Boughattas S. Role of color Doppler in differentiation of Graves' disease and thyroiditis in thyrotoxicosis. *World J Radiol*. 2013;5(4):178-183.
8. Duman E, Aslan A, Buz A, *et al*. Interobserver and intraobserver reliability in sonoelastographic assessment of thyroid nodules. *Ultrasound Q*. 2023;39(1):53-60.
9. Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine learning and deep learning in medical imaging: intelligent imaging. *J Med Imaging Radiat Sci*. 2019;50(4):477-487.
10. Yoon YE, Kim S, Chang HJ. Artificial intelligence and echocardiography. *J Cardiovasc Imaging*. 2021;29(3):193-204.
11. Peng S, Liu Y, Lv W, *et al*. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. *Lancet Digit Health*. 2021;3(4):e250-e259.
12. Lee KS, Park H. Machine learning on thyroid disease: a review. *Front Biosci (Landmark Ed)*. 2022;27(3):101.
13. Zhu YC, Du H, Jiang Q, *et al*. Machine learning assisted Doppler features for enhancing thyroid cancer diagnosis: a multi-cohort study. *J Ultrasound Med*. 2022;41(8):1961-1974.
14. Qiao T, Liu S, Cui Z, *et al*. Deep learning for intelligent diagnosis in thyroid scintigraphy. *J Int Med Res*. 2021;49(1):300060520982842.
15. Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. A review of medical image data augmentation techniques for deep learning applications. *J Med Imaging Radiat Oncol*. 2021;65(5):545-563.
16. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. *Transl Vis Sci Technol*. 2020;9(2):14.
17. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Paper Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016, Las Vegas, NV.
18. Chollet F. Xception: deep learning with depthwise separable convolutions. Paper Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017, Honolulu, HI.
19. Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. Paper Presented at: Proceedings of the 36th International Conference on Machine Learning; June 2019, Long Beach.
20. Lee S-Y, Kang H, Jeong J-H, Kang D-Y. Performance evaluation in [18F]Florbetaben brain PET images classification using 3D Convolutional Neural Network. *Plos One*. 2021;16(10):e0258214.
21. Ota H, Amino N, Morita S, *et al*. Quantitative measurement of thyroid blood flow for differentiation of painless thyroiditis from Graves' disease. *Clin Endocrinol (Oxf)*. 2007;67(1):41-45.
22. Vita R, Di Bari F, Perelli S, Capodicasa G, Benvenga S. Thyroid vascularization is an important ultrasonographic parameter in untreated Graves' disease patients. *J Clin Transl Endocrinol*. 2019;15:65-69.
23. Bayramoglu Z, Kandemirli SG, Akyol Sari ZN, *et al*. Superb microvascular imaging in the evaluation of pediatric Graves disease and hashimoto thyroiditis. *J Ultrasound Med*. 2020;39(5):901-909.
24. Narang A, Bae R, Hong H, *et al*. Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. *JAMA Cardiol*. 2021;6(6):624-632.
25. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med*. 2018;1(1):6.
26. Zhang J, Gajjala S, Agrawal P, *et al*. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation*. 2018;138(16):1623-1635.
27. Kusunose K, Haga A, Inoue M, Fukuda D, Yamada H, Sata M. Clinically feasible and accurate view classification of echocardiographic images using deep learning. *Biomolecules*. 2020;10(5):665.
28. Zhang Q, Zhang S, Pan Y, *et al*. Deep learning to diagnose Hashimoto's thyroiditis from sonographic images. *Nat Commun*. 2022;13(1):3759.
29. Vitti P. Grey scale thyroid ultrasonography in the evaluation of patients with Graves' disease. *Eur J Endocrinol*. 2000;142(1):22-24.
30. Vitti P, Rago T, Mazzeo S, *et al*. Thyroid blood flow evaluation by color-flow Doppler sonography distinguishes Graves' disease from Hashimoto's thyroiditis. *J Endocrinol Invest*. 1995;18(11):857-861.
31. Micucci M, Iula A. Recent advances in machine learning applied to ultrasound imaging. *Electronics (Basel)*. 2022;11(11):1800.
32. Akkus Z, Cai J, Boonrod A, *et al*. A survey of deep-learning applications in ultrasound: artificial intelligence–powered ultrasound for improving clinical workflow. *J Am Coll Radiol*. 2019;16(9, Part B):1318-1328.
33. Moustafa AF, Cary TW, Sultan LR, *et al*. Color Doppler ultrasound improves machine learning diagnosis of breast cancer. *Diagnostics*. 2020;10(9):631.
34. Gómez-Flores W, Pereira W. Gray-to-color image conversion in the classification of breast lesions on ultrasound using pre-trained deep neural networks. *Med Biol Eng Comput*. 2023;61(12):3193-3207.
35. Chen H, Yang B-W, Qian L, *et al*. Deep learning prediction of ovarian malignancy at US compared with O-RADS and expert assessment. *Radiology*. 2022;304(1):106-113.
36. Corona G, Biagini C, Rotondi M, *et al*. Correlation between, clinical, biochemical, color Doppler ultrasound thyroid parameters, and CXCL-10 in autoimmune thyroid diseases. *Endocr J*. 2008;55(2):345-350.
37. Tozzoli R, Bagnasco M, Giavarina D, Bizzaro N. TSH receptor autoantibody immunoassay in patients with graves' disease: improvement of diagnostic accuracy over different generations of methods. Systematic review and meta-analysis. *Autoimmun Rev*. 2012;12(2):107-113.
38. Matthes A, Wolf F, Schmiemann G, Gagyor I, Bleidorn J, Markwart R. Point-of-care laboratory testing in primary care: utilization, limitations and perspectives of general practitioners in Germany. *BMC Prim Care*. 2023;24(1):96.
39. Hamill C, Ellis PK, Johnston PC. Point of care thyroid ultrasound (POCUS) in endocrine outpatients: a pilot study. *Ulster Med J*. 2020;89(1):21-24.
40. Patti RK, Kaur A, Somal N, Dalsania N, Lu T, Kupfer Y. Thyrotoxic periodic paralysis-still a diagnostic challenge. *Proc (Bayl Univ Med Cent)*. 2022;35(6):863-865.
41. Vrijsen BEL, Haitjema S, Westerink J, Hulsbergen-Veelken CAR, van Solinge WW, ten Berg MJ. Shorter laboratory turnaround time is associated with shorter emergency department length of stay: a retrospective cohort study. *BMC Emerg Med*. 2022;22(1):207.
42. Khan AI, Khan M, Khan R. Artificial intelligence in point-of-care testing. *Ann Lab Med*. 2023;43(5):401-407.
43. Mollura DJ, Culp MP, Pollack E, *et al*. Artificial intelligence in low- and middle-income countries: innovating global health radiology. *Radiology*. 2020;297(3):513-520.
44. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6(1):60.

Downloaded from https://academic.oup.com/jcem/article/109/11/2872/7645064 by guest on 18 April 2025

45. Snider EJ, Hernandez-Torres SI, Hennessey R. Using ultrasound image augmentation and ensemble predictions to prevent machine-learning model overfitting. *Diagnostics*. 2023;13(3):417.
46. Rebuffi S-A, Gowal S, Calian DA, Stimberg F, Wiles O, Mann TA. Data augmentation can improve robustness. *Adv Neural Inf Process Syst*. 2021;34:29935-29948.
47. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. *Gigascience*. 2017;6(5):1-9.
48. Fox M, Schoeffmann K. The impact of dataset splits on classification performance in medical videos. Paper Presented at: Proceedings of the 2022 International Conference on Multimedia Retrieval; June 2022, Newark, NJ.
49. Zheng D, Yang Y, Li W. A method of dividing clinical data set for medical image AI training. Paper Presented at: Proceedings Volume 11318, Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications; 2020, Houston, TX.
50. Barinov L, Jairaj A, Becker M, *et al*. Impact of data presentation on physician performance utilizing artificial intelligence-based computer-aided diagnosis and decision support systems. *J Digit Imaging*. 2019;32:408-416.
51. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *Br Med Bull*. 2021;139(1):4-15.
52. Xu J, Wang N, Chu T, Yang B, Jian X, Cui Y. A high-frequency mechanical scanning ultrasound imaging system. *Biosensors*. 2022;13(1):32.
53. Chen Y-L, Chiang HK. Development of single-channel dual-element custom-made ultrasound scanner with miniature optical position tracker for freehand imaging. *Biosensors*. 2023;13(4):431.
54. Alaie M, Tramutola A, Mukamal D. A case report on methimazole-induced severe hypothyroidism. *Cureus*. 2022;14(1):e21339.
55. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019;6(2):94.
56. Polevikov S. Advancing AI in healthcare: a comprehensive review of best practices. *Clinica Chimica Acta*. 2023;548:117519.
57. Kim JA-O, Baek HA-O, Ha J, *et al*. Differential diagnosis of thyrotoxicosis by machine learning models with laboratory findings. *Diagnostics (Basel)*. 2022;12(6):1468.
58. Wiersinga WM, Poppe KG, Effraimidis G. Hyperthyroidism: aetiology, pathogenesis, diagnosis, management, complications, and prognosis. *Lancet Diabetes Endocrinol*. 2023;11(4):282-298.
59. Sharma A, Stan MN. Thyrotoxicosis: diagnosis and management. *Mayo Clin Proc*. 2019;94(6):1048-1064.
60. Bobbia X, Abou-Badra M, Hansel N, *et al*. Electronic address tpgc. Changes in the availability of bedside ultrasound practice in emergency rooms and prehospital settings in France. *Anaesth Crit Care Pain Med*. 2018;37(3):201-205.
61. Cao CL, Li QL, Tong J, *et al*. Artificial intelligence in thyroid ultrasound. *Front Oncol*. 2023;13:1060702.
62. Choi YJ, Baek JH, Park HS, *et al*. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: initial clinical assessment. *Thyroid*. 2017;27(4):546-552.

Downloaded from https://academic.oup.com/jcem/article/109/11/2872/7645064 by guest on 18 April 2025