*Original papers*

*Med Ultrason 2025, Vol. 27, no. 1, 26-31*
**DOI:** 10.11152/mu-4432

# The impact of deep learning on diagnostic performance in the differentiation of benign and malignant thyroid nodules

**Esat Kaba[1], Merve Solak[1], Ayşenur Topçu Varlık[1], Yusuf Çubukçu[1], Lütfullah Sağır[1], Kubilay Muhammed Sünnetci[2,3], Ahmet Alkan[3], Hasan Gündoğdu[4], Fatma Beyazal Çeliker[1], Mehmet Beyazal[1]**

[1]Recep Tayyip Erdogan University, Department of Radiology, Rize, [2]Osmaniye Korkut Ata University, Department of Electrical and Electronics Engineering, Osmaniye, [3]Kahramanmaraş Sütçü İmam University, Department of Electrical and Electronics Engineering, Kahramanmaraş, [4]Samsun University, Department of Radiology, Samsun, Turkey

**Abstract**

**Aims:** This study aims to use deep learning (DL) to classify thyroid nodules as benign and malignant with ultrasonography (US). In addition, this study investigates the impact of DL on the diagnostic success of radiologists with different experiences. **Material and methods:** This study included 576 US images of thyroid nodules. The dataset was divided into 80% training and 20% test sets. Four radiologists with different levels of experience classified the images in the test set as benign-malignant. A DL model was then trained with the train set and predicted benign-malignant for the test set. Then, the output of the DL model for each nodule in the test set was presented to 4 radiologists, who were asked to make a benign-malignant classification again considering these DL results. **Results:** The accuracy of the DL model was 0.9391. The accuracy for junior resident (JR) 1, JR 2, senior resident (SR), and senior radiologist (Srad) before DL-assisting were 0.7043, 0.7826, 0.8435, and 0.8522 respectively. The accuracy in DL-assisted classifications was 0.9130, 0.8696, 0.9304, and 0.9043 for JR 1, JR2, SR, and Srad, respectively. DL assistance changed the decisions of less experienced radiologists more than more experienced radiologists. **Conclusion:** The DL model has superior accuracy in classifying thyroid nodules as benign-malignant with US images than radiologists with different levels of experience. Additionally, all radiologists, and most notably less experienced radiology residents, increased their accuracy in DL-assisted predictions.

**Keywords:** thyroid nodule; ultrasonography; deep learning

## Introduction

Thyroid nodules are lesions that are radiologically and pathologically distinct from normal thyroid parenchyma and may be cystic or solid [1]. Most are asymptomatic and may be discovered incidentally during a neck examination or imaging for other indications. The prevalence of thyroid nodules in the general population ranges from 2% to 65% [2].

Ultrasonography (US) typically serves as the primary imaging modality for evaluating thyroid nodules, and its widespread use, even among healthy populations, has led to an increasing detection of these nodules. The primary advantages of the US are its reliability and accessibility, allowing clinicians to identify and classify thyroid nodules efficiently. However, the US is a user-dependent technique, leading to variability in diagnostic accuracy [3].

Malignant nodules constitute approximately 5–15% of these findings. When there is suspicion of malignancy based on US findings, a fine-needle aspiration biopsy (FNAB) is usually performed [4]. While FNAB is a valu-

Received 21.06.2024 Accepted 08.08.2024
Med Ultrason
2025, Vol. 27, No 1, 26-31
Corresponding author: Esat Kaba
Recep Tayyip Erdogan University,
Department of Radiology,
Rize, 53100,Turkey
E-mail: esatkaba04@gmail.com
Phone/fax: + 90 464 213 04 91
+ 90 464 217 03 64

able diagnostic tool, it is invasive and carries the risk of complications. Furthermore, the nondiagnostic rate of FNAB can vary significantly, ranging from 6.5% to 36.4% across different studies. This variability underscores the importance of accurately identifying nodules that do not exhibit high-risk features to avoid unnecessary invasive procedures, reduce patient anxiety, and manage healthcare resources effectively [5].

To address the limitations of the US and reduce the number of unnecessary biopsies, several reporting systems have been developed. The American College of Radiology created the Thyroid Imaging Reporting and Data System (TI-RADS), which is one of the most widely used systems. It is based on the shape, border, and echogenic focus (calcification) content of nodules, as well as their composition (solid/cystic), and echogenicity (hypoechoic, isoechoic, and hyperechoic). TI-RADS aims to standardize thyroid nodule assessment, determine the risk of malignancy, and guide the need for biopsy, thereby enhancing the overall efficiency and effectiveness of thyroid nodule management [6].

Although TI-RADS standardizes the reporting of thyroid nodules, significant interobserver variability remains its major limitation. Differences in evaluating sonographic features like hypoechogenicity and punctate echogenic foci directly impact the TI-RADS categorization of nodules [7]. Therefore, new methods are needed to facilitate this challenging and time-consuming process and to minimize interobserver differences. For this reason, many recent studies have examined computer-aided diagnosis methods in medical imaging. Deep learning (DL)-based artificial intelligence (AI) models are frequently used in these studies. DL is a network that uses interconnected neuronal layers to build an output model based on the validated diagnoses of the data. DL has made significant advances in interpretation, segmentation, localization, and classification in medical imaging analysis with the development of convolutional neural networks [8].

In this context, this study aims to test the accuracy and effectiveness of DL-based AI models in differentiating between benign and malignant thyroid nodules. Additionally, an important aspect of this study is to determine how these DL-based models affect the decision-making of radiologists with varying levels of experience. Ultimately, our study aims to demonstrate the potential of AI-assisted diagnostic systems in the clinical application of thyroid nodules.

## Material and methods

This study was designed retrospectively, and the data used in this study were not used in any other study.

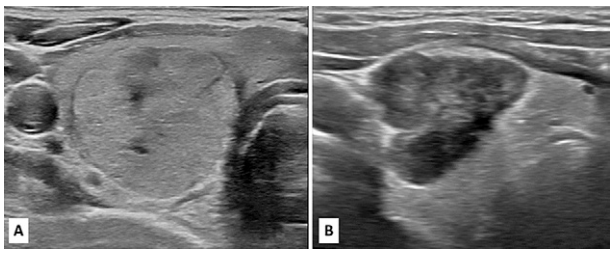


**Fig 1.** Benign thyroid nodule (A), Malignant thyroid nodule (B).

### *Patient selection*

Patients who underwent thyroid FNAB in the interventional radiology unit of our hospital between March 2022 and October 2022 were included in this study. The inclusion criteria comprised patients who had ultrasound images taken immediately before FNAB, with all images obtained using the same device and the same probe. The exclusion criteria were as follows: patients with unclear pathological results, non-diagnostic pathological results (Bethesda I), poor image quality, or biopsy results indicating atypia of undetermined significance (Bethesda III) or suspicion of follicular neoplasia (Bethesda IV) who did not undergo surgery. During the period of this study, a total of 308 patients underwent thyroid FNAB. After applying the exclusion criteria, the final dataset consists of 576 images (481 benign, 95 malignant) from a total of 178 patients. All images were obtained in the transverse plane with the same device (Philips Affiniti 70G) and the same probe (Philips eL18-4 (22-2 MHz)). Patient information was anonymized for standardization of images and images were saved in standard sizes (1920x1080x3). A 4-year radiology resident and a 10-year interventional radiologist performed all US and FNAB procedures together. On average, 2-3 ultrasound images were obtained along the largest dimension of the thyroid nodule. The pathological examination of the nodules was conducted by a pathologist experienced in thyroid nodules, following the Bethesda classification system, and all biopsy results were collected from the hospital system. The US image of a sample benign and malignant thyroid nodule in the dataset is shown in figure 1.

### *Deep Learning*

In this study, a DL model is used to extract image features, and a different architecture is used to classify the extracted features. It can be stated here that a new model has been created by integrating these two different architectures used within the scope of the study. Features of US images of thyroid nodules are extracted using AlexNet DL architecture [9]. Herein, they are obtained from AlexNet's last fully connected layer. The number of image features obtained for each image is equal to 1000. Afterward, these images are split into 80% (461 images)

and 20% (115 images) training and test sets, respectively. This split is randomized for all classes. That is, the number of images used in the training phase for the dataset is 461. In addition, the data used in the study is trained with the help of the Fine KNN classifier using the 10-fold cross-validation method, as it is mostly used in the literature [10]. For the KNN classifier used in the study, distance metric and distance weight are Euclidean and Equal, respectively. Based on the trained AlexNet+Fine KNN architecture, the model performance is calculated for the 20% test set (115 images). By using these models together, confusion matrices are obtained for this integrated model. Afterward, classification performance metrics are calculated from the matrices obtained for this total model, and it is tested so that the models work successfully. In this study, accuracy, sensitivity, specificity, precision, and F1 scores are used as performance metrics. The reason for using various performance metrics in this study is to prove the reliability of the developed model by testing it from different aspects, and they are widely used in such studies.

***Radiologist evaluation***

In our study, five radiologists evaluated US test set images (115 images). Three junior radiology residents (JR) with 1 year of experience, a senior radiology resident (SR) with 4 years of experience, and a head and neck radiologist (Srad) with 15 years of experience interpreted the test set images. All participants were blinded to the pathology results. Except for one JR, the other four participants first predicted benign-malignant for all images in the test set. Then, the result provided by DL for each nodule in the test set was shared with the radiologists and they were asked to re-evaluate using the DL results. Another JR made a decision directly with the DL assistance. The reason for this was to simulate the diagnostic application we want to develop in the future and to validate the results of other JRs. Performance metrics were obtained for all participants before and after DL assistance.

## Results

Of the 178 patients in the dataset, 139 were female and 39 were male. The mean age of the patients was 56.9 years. The mean transverse diameter of the nodules was 20.6 mm, and the mean anteroposterior diameter was 15.5 mm. The accuracy, sensitivity, and specificity results achieved using the AlexNet+ Fine KNN model on a test set of 115 images are 0.9391, 0.9688, and 0.7895, respectively. Before implementing DL assistance, SRad exhibited the highest accuracy among radiologists, with an accuracy value of 0.8522, followed by SR with an accuracy value of 0.8435. These findings demonstrate that the DL model surpasses all radiologists in diagnostic accuracy.

Table I. Performance metrics of radiologist and Deep Learning

| Dataset 1 | Accuracy | Sensitivity | Specificity | Precision | $F_1$ Score |
|---|---|---|---|---|---|
| JR 1 | 0.7043 | 0.7188 | 0.6316 | 0.9079 | 0.8023 |
| JR 2 | 0.7826 | 0.8125 | 0.6316 | 0.9176 | 0.8619 |
| SR | 0.8435 | 0.9278 | 0.3889 | 0.8911 | 0.9091 |
| SRad | 0.8522 | 0.8854 | 0.6842 | 0.9341 | 0.9091 |
| DL | 0.9391 | 0.9688 | 0.7895 | 0.9588 | 0.9637 |

JR: Junior Resident, SR: Senior Resident, SRad: Senior Radiologist

Table II. Performance metrics of radiologist with Deep Learning-assisting

| Dataset 1 | Accuracy | Sensitivity | Specificity | Precision | $F_1$ Score |
|---|---|---|---|---|---|
| JR 1 | 0.9130 | 0.9688 | 0.6316 | 0.9300 | 0.9490 |
| JR 2 | 0.8696 | 0.9271 | 0.5789 | 0.9175 | 0.9223 |
| JR 3 | 0.8696 | 0.8958 | 0.7368 | 0.9451 | 0.9198 |
| SR | 0.9304 | 0.9792 | 0.6842 | 0.9400 | 0.9592 |
| SRad | 0.9043 | 0.9479 | 0.6842 | 0.9381 | 0.9430 |

JR: Junior Resident, SR: Senior Resident, SRad: Senior Radiologist

Deep learning assistance improved the diagnostic performance of all radiologists to varying degrees. The DL-assisted accuracy results are 0.9130, 0.8696, 0.8696, 0.9304, and 0.9043 for JR 1, JR 2, JR 3, SR, and Srad, respectively. The sensitivity is 0.9688, 0.9271, 0.8958, 0.9792 and 0.9479 respectively. DL assistance improved the accuracy of less experienced radiologists more than experienced radiologists. In other words, while JR 1 had an accuracy of 0.7043 before DL assistance, it increased to 0.9130 accuracy with DL assistance. On the other hand, the SR has an accuracy value of 0.8435 and 0.9304, respectively, without and with DL assistance. Detailed results are given in Table I and Table II.

The confusion matrix of the validation and test set results obtained from the AlexNet+Fine KNN model is given in figure 2.

## Discussion

In this study, we conducted the potential impact of DL on the diagnostic accuracy of both inexperienced and experienced radiologists. Our findings indicate that the DL model we employed significantly enhanced the performance of inexperienced radiologists and also contributed to the performance of experienced radiologists.

Specifically, JR 1 demonstrated a diagnostic accuracy rate of 0.7043 without DL assistance, which increased to 0.9130 with DL assistance. Similarly, JR 2's accuracy improved from 0.7826 without DL assistance to 0.8696 with DL assistance. JR 3, possessing the same level of experience as JR 1 and JR 2, achieved an accuracy rate of 0.8696 with DL assistance. Conversely, SR's accuracy increased from 0.8435 before DL assistance to 0.9304 with DL assistance. When these results are evaluated, it can be seen that DL-based results changed the decisions of less experienced radiologists more than more experienced radiologists. Thus, the accuracy rate of less experienced radiologists increased significantly.

Thyroid nodules are a common problem in thyroid diseases [11,12]. Epidemiologic studies have shown that the prevalence of palpable thyroid nodules is approximately 1-5%, even in people living in iodine-sufficient regions of the world [12]. However, with the increasing use of high-resolution US, the most effective imaging modality for thyroid nodule screening, thyroid nodules are being detected at higher prevalences in randomly selected individuals [13-15]. With this effective screening method, the probability of newly diagnosed cases of thyroid neoplasms from thyroid nodules has exceeded the estimate in the American Thyroid Association guidelines published in 2009 [12,16,17]. However, this increase in patient diagnosis has made it difficult to clinically optimize long-term health outcomes.

Although many studies have shown that the US alone is useful in differentiating benign and malignant nodules, significant interobserver variability in the identification of these features limits diagnostic accuracy [18,19]. We recognize that diagnostic success is dependent on many factors, especially the significant challenges in the clinical practice setting. In this case, clinical approaches such as neck US, FNAB cytology, or serial follow-up may be applied [19]. However, as noted earlier, due to the high prevalence of nodules in the population, performing FNAB without proper biopsy indications can lead to a substantial number of patients receiving non-diagnostic or inconclusive results [20].

To overcome the diversity of diagnostic methods and approaches in the increasing prevalence of thyroid nodular diseases, recent studies in this field have tried to provide accurate identification of US data using AI. In early studies using AI, results have shown that it can distinguish malignant nodules from benign nodules with an accuracy similar to expert radiologists [21,22]. Our study results also demonstrate that DL-based AI models can exhibit better diagnostic performance, particularly compared to less experienced radiologists, and can assist inexperienced radiologists in the diagnostic decision-making process. Additionally, successful DL models hold strong potential for reducing unnecessary biopsy rates and, consequently, patient anxiety by non-invasively predicting benign or malignant conditions.

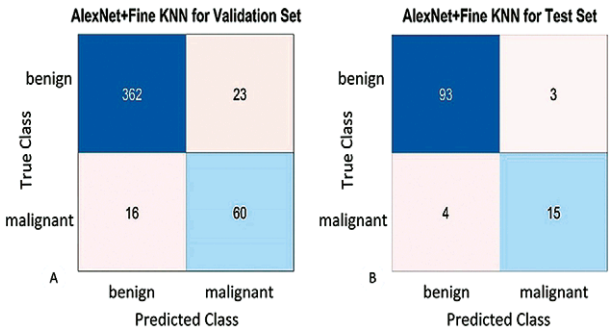


**Fig 2.** Confusion matrix of validation set (A), Confusion matrix of test set (B).

In recent years, several DL-based studies have been conducted. Yi-Cheng Zhu et al using the VGG-16 architecture, reported that the DL model produced more successful results from radiologists with different experiences [23]. In addition, Sui Peng et al used different DL architectures in their study and demonstrated the significant contribution of these architectures to radiologist performance [24]. In another study, Xi Wei et al used four different DL models and an ensemble model based on 26541 thyroid ultrasound images. The ensemble model showed a superior performance with an accuracy of 93.70%. Also in their study, three radiologists with different levels of experience had 91.55%, 87.26%, and 93.07% accuracy respectively. Additionally, with the DL assistance of the ensemble model, the accuracy of the radiologists increased to 96.54% [25]. These findings, along with the results of our study underscore that leveraging AI in thyroid disease ultrasound scanning and diagnosis not only enhances diagnostic accuracy for radiologists but also significantly alleviates their workload by non-invasively and effectively distinguishing between benign and malignant nodules [26].

This retrospective study had some limitations. Firstly, the number of patients was small and the numbers of patients in benign and malignant groups were different. Secondly, excluding patients in Bethesda groups III and IV who did not undergo surgery was a limitation, as it affected the number of patients in the benign and malignant groups in the dataset. Thirdly, radiologists evaluated thyroid nodules from a single slice US image.Lastly image features were extracted with a single DL architecture. In the future, benign-malignant classification studies can be performed on three-dimensional images or US videos using different DL architectures with more balanced datasets containing a larger number of patients.

## Conclusion

In our study, we investigated the success of DL in the benign-malignant classification of thyroid nodules and its impact on radiologists with different levels of experience. According to our results, the DL model has superior accuracy in classifying thyroid nodules as benign-malignant with US images than radiologists with different levels of experience. Additionally, all radiologists, and most notably less experienced radiology residents, increased their accuracy in differentiating benign-malignant thyroid nodules with DL-assisting. Considering the high incidence of thyroid nodules, DL can assist all radiologists with a high accuracy rate and support the practice of radiology for thyroid nodules. In other words, while DL reduces the workload of all radiologists, especially experienced radiologists, it can also become a part of radiology education by contributing positively to the success of inexperienced radiologists.

**Conflict of interest:** none

## References

1. Youssef A, Abd-Elmonem MH, Ghazy RAM, El Shafei MM, Zahran M. The diagnostic value of ultrasonography in detection of different types of thyroid nodules. Egypt J Otolaryngol 2020;36:23.
2. Mu C, Ming X, Tian Y, et al. Mapping global epidemiology of thyroid nodules among general population: A systematic review and meta-analysis. Front Oncol 2022;12:1029926.
3. Shin JH, Baek JH, Chung J, et al. Ultrasonography diagnosis and imaging-based management of thyroid nodules: Revised Korean society of thyroid radiology consensus statement and recommendations. Korean J Radiol 2016;17:370–395.
4. Lee YH, Baek JH, Jung SL, Kwak JY, Kim JH, Shin JH. Ultrasound-guided fine needle aspiration of thyroid nodules: A consensus statement by the Korean society of thyroid radiology. Korean J Radiol 2015;16:391–401.
5. Ziemiańska K, Kopczyński J, Kowalska A. Repeated non-diagnostic result of thyroid fine-needle aspiration biopsy. Contemp Oncol (Pozn) 2016;20:491-495.
6. Tessler FN, Middleton WD, Grant EG, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. J Am Coll Radiol 2017;14:587–595.
7. Hoang JK, Middleton WD, Tessler FN. Update on ACR TI-RADS: Successes, Challenges, and Future Directions, From the AJR Special Series on Radiology Reporting and Data Systems. AJR Am J Roentgenol 2021;216:570-578.
8. Sharifi Y, Bakhshali MA, Dehghani T, DanaiAshgzari M, Sargolzaei M, Eslami S. Deep learning on ultrasound images of thyroid nodules. Biocybern Biomed Eng 2021;41:636–655.
9. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM 2017;60:84-90.
10. Sunnetci KM, Alkan A. Biphasic majority voting-based comparative COVID-19 diagnosis using chest X-ray images. Expert Syst Appl 2023;216:119430.
11. Popoveniuc G, Jonklaas J. Thyroid Nodules. Med Clin North Am 2012;96:329–349.
12. Haugen BR, Alexander EK, Bible KC, 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.
13. Tan GH, Gharib H. Thyroid incidentalomas: Management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. Ann Intern Med 1997;126:226–231.
14. Tong WJ, Wu SH, Cheng MQ, et al. Integration of Artificial Intelligence Decision Aids to Reduce Workload and Enhance Efficiency in Thyroid Nodule Management. JAMA Netw Open 2023;6:e2313674.
15. Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ, Cooper DS. The diagnosis and management of thyroid nodules: a review. JAMA 2018;319:919–924.
16. Davies L, Welch HG. Current thyroid cancer trends in the United States. JAMA Otolaryngol Head Neck Surg 2014;140:317–322.
17. Leennhardt L, Bernier MO, Boin-Pineau MH, et al. Advances in diagnostic practices affect thyroid cancer incidence in France. Eur J Endocrinol 2004;150:133–139.
18. Fresilli D, Grani G, De Pascali ML, et al. Computer-aided diagnostic system for thyroid nodule sonographic evaluation outperforms the specificity of less experienced examiners. J Ultrasound 2020;23:169–174.
19. Ram N, Hafeez S, Qamar S, et al. Diagnostic validity of ultrasonography in thyroid nodules. J Pak Med Assoc 2015;65:875–878.
20. Nabahati M, Moazezi Z, Fartookzadeh S, Mehraeen R, Ghaemian N, Sharbatdaran M. The comparison of accuracy of ultrasonographic features versus ultrasound-guided fine-needle aspiration cytology in diagnosis of malignant thyroid nodules. J Ultrasound 2019;22:315–321.
21. 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:InitialClinicalAssessment.Thyroid2017;27:546–552.
22. Gitto S, Grassi G, De Angelis C, et al. A computer-aided diagnosis system for the assessment and characterization of low-to-high suspicion thyroid nodules on ultrasound. Radiol Med 2019;124:118–125.
23. Zhu YC, Jin PF, Bao J, Jiang Q, Wang X. Thyroid ultrasound image classification using a convolutional neural network. Ann Transl Med 2021;9:1526.
24. 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:e250–e259.
25. Wei X, Gao M, Yu R, et al. Ensemble Deep Learning Model for Multicenter Classification of Thyroid Nodules on Ultrasound Images. Med Sci Monit 2020;26: e926096.
26. Cao CL, Li QL, Tong J, et al. Artificial intelligence in thyroid ultrasound. Front Oncol 2023;13:1060702.