Research paper

Cite as: Żyłka A, Rafało M, Dedecjus M, Dobruch-Sobczak KS: Application of artificial intelligence in the ultrasonographic diagnosis of thyroid nodules. J Ultrason 2025; 25: 22. doi: 10.15557/JoU.2025.0022.

Submitted: 21.06.2025
Accepted: 21.08.2025
Published: 30.09.2025

Application of artificial intelligence in the ultrasonographic diagnosis of thyroid nodules

Agnieszka Żyłka[1], Mariusz Rafało[2], Marek Dedecjus[1], Katarzyna Sylwia Dobruch-Sobczak[3]

[1] *Department of Endocrine Oncology and Nuclear Medicine, Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland*
[2] *Institute of Information Systems and Digital Economy, Warsaw School of Economics, Poland*
[3] *Radiology Department II, Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland*

Corresponding author: Agnieszka Żyłka; e-mail: agnieszka.zylka.edu@gmail.com

DOI: 10.15557/JoU.2025.0022

Keywords
thyroid nodules; artificial intelligence; convolutional neural network; ultrasound; thyroid cancer

Abstract

Introduction: The aim of this study was to assess whether integrating selected ultrasound features into a convolutional neural network improves its ability to differentiate benign from malignant thyroid nodules. **Material and methods:** A total of 242 patients (196 women, 46 men) with thyroid lesions were included in the study. All patients underwent surgical treatment and histopathological analysis. Thyroid ultrasonography was also performed for all participants. Images were recorded in DICOM and AVI formats, and archived in a local database. Thyroid lesions were assessed according to the EU-TIRADS classification. Convolutional neural network models were developed using established architectures, including DenseNet and VGG16, as well as custom-designed models tailored to the dataset. Hybrid models were created by incorporating selected ultrasound features into these architectures as additional inputs. Performance was compared between the baseline convolutional neural network models and their feature-supported hybrid counterparts. **Results:** Model performance was evaluated using several metrics, including sensitivity and area under the ROC curve. Baseline convolutional neural network models served as the reference, while hybrid variants included structured ultrasound features. The VGG model showed a sensitivity of 0.78, and DenseNet achieved a sensitivity of 0.80 with an AUC of 0.84, demonstrating low variability. Inception models performed similarly, with balanced positive predictive value (PPV) (0.83) and negative predictive value (NPV) (0.74). Custom models also reached AUC values over 0.80. Selected ultrasonography features improved AUC by up to 7%, with additional gains in sensitivity and NPV. **Conclusions:** Eight baseline convolutional neural network models used to differentiate benign from malignant thyroid nodules were enhanced by incorporating five expert-assessed ultrasound features. This hybrid approach improved classification performance across all models, yielding an average AUC increase of approximately 7%.

Introduction

The incidence of thyroid cancer has been steadily increasing in many countries, including Poland. According to the National Cancer Registry, a total of 5,012 new cases were reported in 2022, of which 4,107 were among women[(1)]. The key diagnostic procedures continue to be high-resolution ultrasonography and fine-needle aspiration biopsy[(2)].

In recent years, multiple ultrasound-based risk stratification systems have been proposed, including ACR-TIRADS, (American College of Radiology Thyroid Imaging Reporting and Data System), EU-TIRADS (European Thyroid Imaging Reporting and Data System), and K-TIRADS (Korean Thyroid Imaging Reporting and Data System), all designed to improve malignancy risk assessment of thyroid nodules[(3–5)]. These systems have contributed to improved diagnostic accuracy; however, notable interobserver variability persists, particularly in determining which nodules should undergo biopsy based on imaging features and size criteria[(6)].

A meta-analysis comparing EU-TIRADS, ACR-TIRADS, and K-TIRADS reported sensitivities ranging from 0.68 to 0.82, and specificities between 0.71 and 0.81. The analysis also included S-Detect, a commercially available system based on K-TIRADS that classifies nodules as probably benign or probably malignant. Its sensitivity and specificity (0.73 and 0.78, respectively) did not differ significantly from those of classifications performed by radiologists[(7)]. S-Detect has demonstrated diagnostic utility for both junior and experienced radiologists.

Several studies and meta-analyses have also evaluated custom software based on machine learning and convolutional neural networks, reporting diagnostic performance comparable to that of expert assessment, with area under the curve values ranging from 0.76 to 0.98[8].

Convolutional Neural Networks (CNNs) are a class of deep learning models designed to automatically extract features from multidimensional data such as images, sound, or video. A CNN is composed of multiple layers that progressively extract and learn hierarchical features from input images. Typical architectures include convolutional layers for feature detection, pooling layers for spatial downsampling, dropout layers to reduce overfitting, and flatten layers that convert multidimensional outputs into a format suitable for fully connected classification layers. This layered structure enables CNNs to capture both low-level textures and high-level semantic patterns essential for accurate image classification. A diagram of the CNN network, indicating individual components, is presented in Fig. 1. CNN models are widely used in medical imaging tasks such as breast cancer classification, thyroid nodules examination, or skin cancer detection[9–12].

Despite the promising diagnostic performance of CNNs in thyroid cancer classification based on ultrasound imaging, these models typically operate in isolation from structured clinical input. In routine practice, diagnosticians evaluate a range of sonographic features, such as echogenicity, margins, and calcifications, which are not always explicitly utilized by data-driven models.

The aim of this study was to evaluate whether incorporating selected ultrasound features, as assessed by experienced diagnosticians, into a CNN could improve diagnostic performance. It was hypothesized that combining image-based deep learning with structured semantic features would enhance model accuracy in differentiating malignant from benign thyroid nodules.

Materials and methods

A total of 242 patients (196 women, 46 men) with focal thyroid lesions were included in the study. All patients (n = 242) underwent surgical treatment at the Department of Oncological Endocrinology and Nuclear Medicine, National Institute of Oncology, Warsaw. The study was approved by the local institutional review board (decision number: 22/2024). Participants met the following inclusion criteria:

- age ≥18 years;
- presence of a focal lesion within the thyroid;
- qualification for surgical treatment based on fine-needle aspiration biopsy (FNAB) classified as Bethesda category III–VI; or
- for Bethesda category II, the presence of clinical symptoms such as dysphagia or dyspnea.

No focal thyroid lesion were excluded.

Ultrasound examination

All patients underwent thyroid ultrasound using a premium-class system (Philips EpiQ 5, Bothell, DC, USA) equipped with a linear transducer (eL18-4, frequency range 2–22 MHz). Images were recorded in Digital Imaging and Communications in Medicine (DICOM) and Audio Video Interleave (AVI) formats and archived in a local database.

Morphological features of the focal lesions were evaluated according to the EU-TIRADS classification. The following sonographic parameters were assessed: echogenicity, echotexture, shape, margins, presence of microcalcifications, vascularity, and composition.

Ultrasound examinations were performed by two experienced ultrasonographers, each with over 10 years of experience in thyroid diagnostics. Initially, images were reviewed independently. Subsequently, all cases were re-evaluated jointly. In approximately 10% of discordant cases, diagnostic consensus was achieved through discussion.

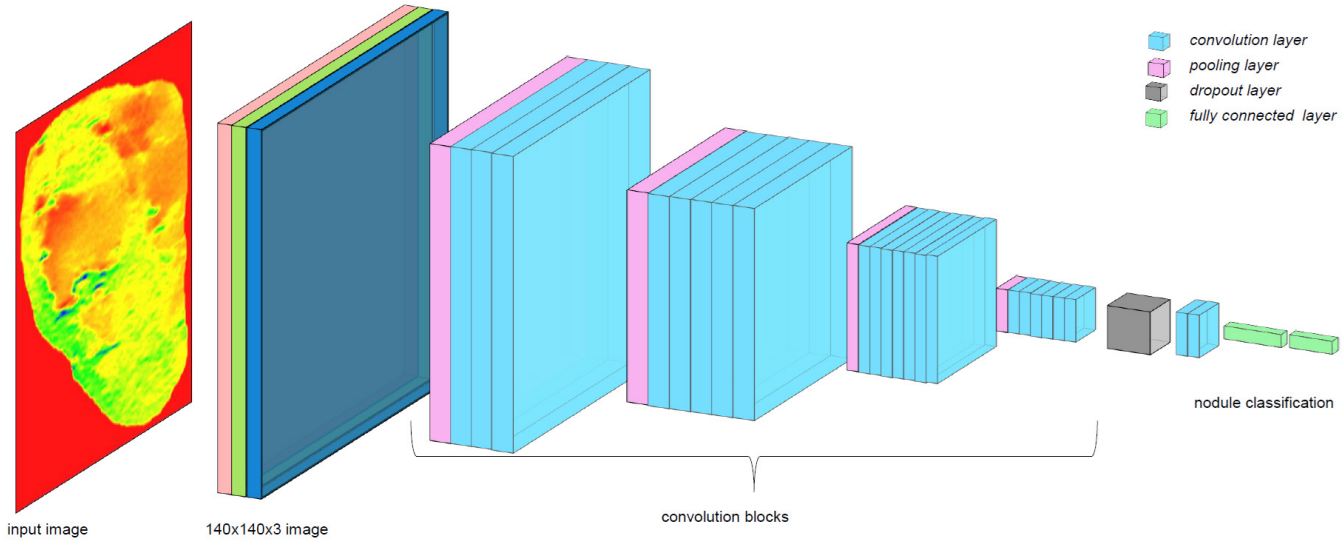

Fig. 1. CNN model building blocks. Source: own study

Patient characteristics

Postoperative histopathological analysis was performed on all 242 focal lesions. Of these, 139 were malignant (113 women, 26 men; mean age: 50 ± 14 years), and 103 were benign (83 women, 20 men; mean age: 48 ± 14 years).

Patients with malignant tumors underwent surgical treatment, either total thyroidectomy or lobectomy with isthmectomy. Central neck compartment lymphadenectomy was performed where clinically indicated. The extent of surgery was determined by the oncological surgeon based on ultrasound findings, FNAB results, and patient symptoms. In the malignant group, EU-TIRADS category 5 was predominant (n = 118; 85%), with fewer lesions classified as EU-TIRADS 4 (n = 14; 10%) and EU-TIRADS 3 (n = 7; 5%). In the benign group, the distribution was more balanced: EU-TIRADS 5 – n = 33 (32%), EU-TIRADS 4 – n = 36 (35%), EU-TIRADS 3 – n = 34 (33%).

Approach

In this study, we employed commonly used CNN architectures and also designed our own CNN models. We verified predefined architectures, such as Visual Geometry Group (VGG), which uses deep stacks of convolutional layers with small filters to capture fine-grained image details, DenseNet (Densely Connected Convolutional Network), which enhances feature propagation by connecting each layer to every other layer, reducing redundancy and improving gradient flow, and Deep Inception Dense CNN (DIDC), which incorporates the inception mechanism. Moreover, we developed three deep networks from scratch (marked as cnn1, cnn2, and cnn3). These custom models differ in the number of convolutional blocks and the number and size of convolutional layers within each block.

Furthermore, to enhance classification performance, we extended the CNN models by incorporating additional binary ultrasound features provided by a medical expert. A total of 29 features, representing clinically relevant characteristics of thyroid nodules, were combined with the image-based outputs of the CNN. These features included echogenicity, nodule shape, margin characteristics, vascularity, composition, presence of calcifications, and other parameters commonly used in thyroid risk stratification systems.

The extended hybrid models were implemented by extending the CNN architecture with a parallel branch that processes the selected tabular features and merges them with the CNN output prior to the final classification layer.

To determine the quality of models, we evaluated sensitivity, specificity (true negative rate (TNR)), positive predictive value (PPV), negative predictive value (NPV), and area under the curve (AUC) metrics. For each metric, we calculated the mean, standard deviation, and 95% confidence intervals. We used 20 iterations for the cross-validation of each model, using a random sampling method[13].

Data preparation process

We utilized 479 images (278 malignant and 201 benign) from 242 patients. For each image, we manually selected the region of interest (the lesion area). Then we extracted this part from the full ultrasound image (Fig. 2). The annotation was performed by an experienced radiologist. All images were resized to 140×140 pixels to ensure compatibility with the input requirements of the specific CNN.

The dataset was split into 75% for training,15% for validation, and 10% for testing model performance.

Results

Base models

The performance metrics for all tested CNN models are listed in Table 1. The table presents the baseline models that serve as reference points, as well as the extended (hybrid) models (marked as H), used for evaluating the impact of integrating structured clinical features with the CNN models.

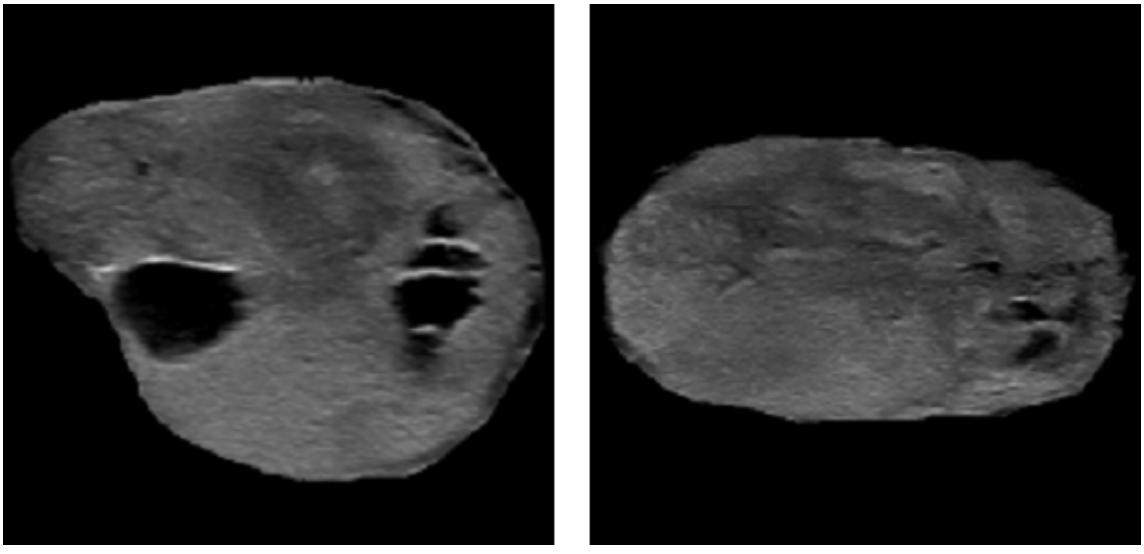

Fig. 2. Sample US images of a thyroid nodule from two planes. Source: own study

Tab. 1. Summary of model results. Models using US features are marked as H (hybrid). Source: own study

Model	Recall	Recall CI	PPV	PPV CI	NPV	NPV CI	TNR	TNR CI
VGG16	0.78 ± 0.12	(0.64, 092)	0.75 ± 0.06	(0.59, 0.91)	0.70 ± 0.09	(0.50, 0.89)	0.63 ± 0.09	(0.43, 0.84)
VGG16 (H)	0.85 ± 0.10	(0.73, 096)	0.80 ± 0.05	(0.72, 0.96)	0.77 ± 0.13	(0.63, 0.94)	0.68 ± 0.06	(0.55, 0.85)
denseNet121	0.80 ± 0.06	(0.55, 0.87)	0.88 ± 0.07	(0.68, 0.96)	0.73 ± 0.05	(0.54, 0.92)	0.83 ± 0.08	(0.66, 0.98)
denseNet121 (H)	0.89 ± 0.07	(0.75, 093)	0.92 ± 0.08	(0.77, 0.98)	0.75 ± 0.07	(0.65, 0.86)	0.84 ± 0.11	(0.71, 0.96)
DIDC	0.80 ± 0.11	(0.56, 0.88)	0.83 ± 0.07	(0.64, 0.94)	0.74 ± 0.11	(0.55, 0.92)	0.77 ± 0.13	(0.59, 0.95)
DIDC (H)	0.85 ± 0.07	(0.62, 0.89)	0.89 ± 0.08	(0.69, 0.97)	0.77 ± 0.05	(0.61, 0.93)	0.81 ± 0.07	(0.65, 0.94)
cnn1	0.80 ± 0.09	(0.58, 0.90)	0.81 ± 0.08	(0.63, 0.94)	0.75 ± 0.08	(0.56, 0.93)	0.73 ± 0.08	(0.55, 0.92)
cnn1 (H)	0.85 ± 0.11	(0.73, 0.96)	0.85 ± 0.08	(0.72, 0.96)	0.79 ± 0.14	(0.63, 0.94)	0.78 ± 0.06	(0.61, 0.94)
cnn2	0.74 ± 0.07	(0.57, 0.89)	0.87 ± 0.06	(0.69, 0.97)	0.68 ± 0.07	(0.49, 0.87)	0.82 ± 0.09	(0.65, 0.97)
cnn2 (H)	0.78 ± 0.10	(0.64, 0.92)	0.92 ± 0.07	(0.82, 1.00)	0.74 ± 0.09	(0.56, 0.90)	0.89 ± 0.08	(0.77, 1.0)
cnn3	0.76 ± 0.09	(0.59, 0.89)	0.88 ± 0.07	(0.70, 0.96)	0.70 ± 0.11	(0.51, 0.87)	0.83 ± 0.07	(0.67, 0.96)
cnn3 (H)	0.85 ± 0.08	(0.73, 0.97)	0.92 ± 0.03	(0.83, 1.00)	0.79 ± 0.06	(0.61, 0.96)	0.87 ± 0.06	(0.73, 1.0)

Among the predefined base model architectures, VGG class model achieved around 0.78 sensitivity, indicating strong detection of malignant cases, but a relatively low NPV (0.70), suggesting that it may struggle to correctly identify benign cases. The DenseNet model achieved 0.80 sensitivity with relatively low variance (0.06). An AUC of 0.84 reflects reliable classification performance. The inception architecture (DIDC) exhibits strong sensitivity (0.80), with balanced PPV (0.83) and NPV (0.74). Our custom models also showed consistent performance with good AUC scores (over 0.80).

Extended hybrid models

Table 2 presents AUC improvement rates achieved when using the hybrid models in comparison with the base models.

Tab. 2. Comparison between base and hybrid models. Source: own study

	AUC		AUC improvement
	Base model	Hybrid model	
VGG16	0.73 ± 0.03	0.81 ± 0.05	11%
denseNet121	0.84 ± 0.04	0.89 ± 0.06	6%
DIDC	0.80 ± 0.06	0.84 ± 0.04	5%
cnn1	0.80 ± 0.02	0.84 ± 0.01	5%
cnn2	0.81 ± 0.02	0.87 ± 0.02	7%
cnn3	0.83 ± 0.03	0.89 ± 0.02	7%

The hybrid models include US features. The Pearson–Matthews correlation coefficient between these features is presented, including the dependencies on the TIRADS category and the occurrence of thyroid cancer (Fig. 3).

Color intensity represents correlation strength, with red indicating positive and blue indicating negative correlations. The cancer variable reveals the strongest correlations with specific nodule features.

To identify the most informative features, we computed feature importance scores. The calculation was performed using a random forest algorithm. Importance was computed as the mean of accumulation of the impurity decrease within each tree in random forest (Mean Decrease Impurity – MDI measure). Based on this analysis, the top five most important features were selected and used as auxiliary inputs to the CNN models (Fig. 4). These included microlobulated margins, deeply hypoechoic echogenicity, ill-defined margins, isoechogenic appearance, and the presence of microcalcifications. The presence (or absence) of these characteristics is commonly recognized by clinicians as indicators of malignancy risk in thyroid nodules. The frequency of occurrence of these features, divided into malignant and benign nodules, is also presented (Tab. 3).

The hybrid models were implemented by extending the CNN architecture with a parallel branch that processes the selected tabular features and merges them with the CNN output before the final classification layer. Incorporating the selected ultrasound features led to consistent improvements across all model architectures. On average (for all models), sensitivity, PPV, specificity, and AUC increased by 7% (each measure).

Discussion

Our study reinforces the findings of previous research demonstrating that deep learning algorithms can achieve reliable diagnostic performance for relatively small sample sizes. In this study, we compared the quality of CNN models with hybrid models supported by selected US features. The highest sensitivity (0.89) was observed in the hybrid DenseNet model, suggesting its value in minimizing false negatives. The cnn3(H) model demonstrated balanced performance (sensitivity 0.85, PPV 0.92, NPV 0.79, AUC 0.89). Most of our models exceeded the AUC threshold of 0.8, suggesting robustness and operational stability.

For context, Lee *et al.* reported an AUC of up to 0.90 based on 5,575 images, using VGG models and a transfer learning approach, while Li *et al.* obtained an AUC ranging from 0.90 to 0.95, depending on the dataset[14–15]. Despite a smaller sample size, our models yielded comparable diagnostic performance.

Moreover, there are studies proving that AI systems may exceed expert-level sensitivity. Lai *et al.* showed that the InceptionV3,

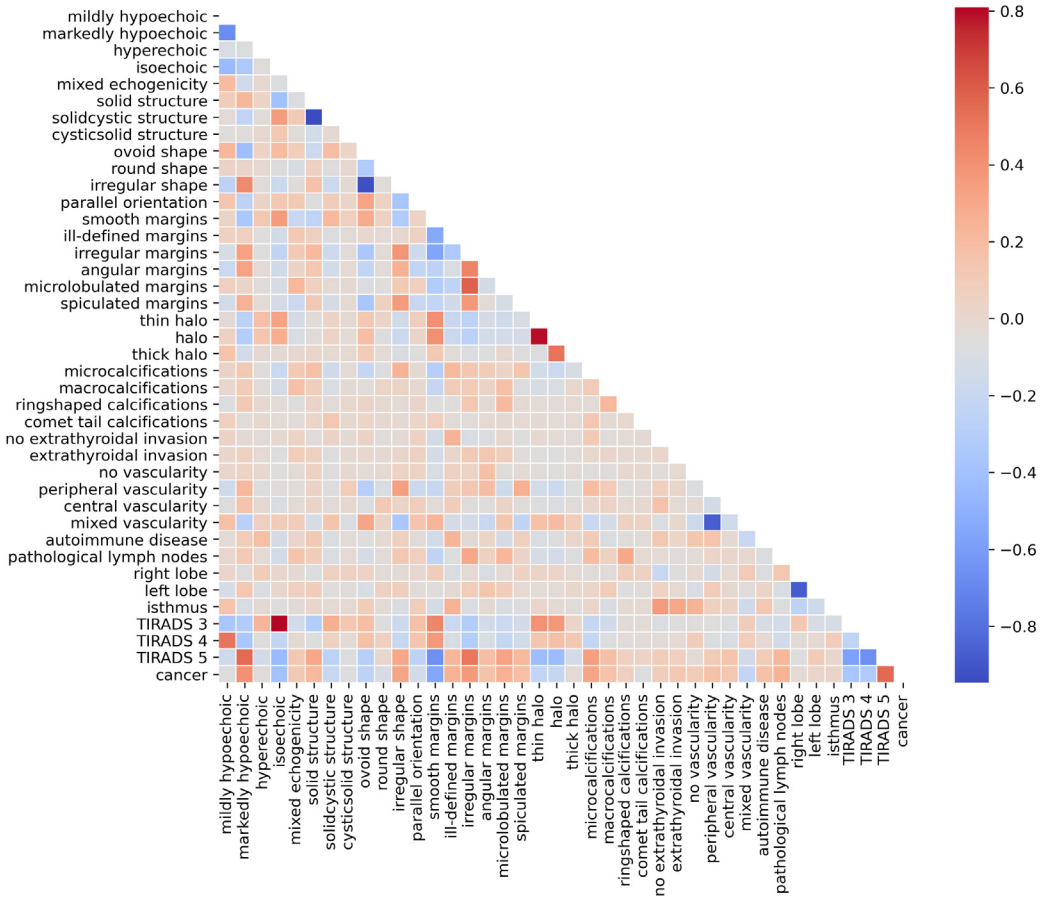

Fig. 3. Correlation matrix for US features. Source: own study

Tab. 3. Number of nodules with specific features. Source: own study

Feature	All	Malignant	Benign
Microlobulated margins	33	28	5
Markedly hypoechoic	85	72	13
Isoechoic	45	7	38
Ill-defined margins	74	57	17
Microcalcifications	46	42	4

DenseNet121, and ResNet50 models achieved AUC values between 0.85 and 0.86, surpassing a senior radiologist (AUC 0.82), while their data fusion model reached an AUC of 0.88[16]. A similar approach was reported by Peng *et al.*, who showed that the ThyNet model achieved an AUC of 0.92, compared to radiologists (0.84; p <0.0001), and its implementation reduced unnecessary biopsies from 61.9% to 35.2%, while the proportion of missed malignancies decreased from 18.9% to 17.0%[17]. In the study by Namsena *et al.*, the CNN system achieved a sensitivity of 0.80 versus 0.40 for the radiologist (p = 0.043), with comparable specificity[18]. In our cohort, the cnn3 model (AUC 0.85, PPV 0.92) also demonstrated potential clinical utility as a decision-support tool.

From a clinical perspective, AI should be viewed as an adjunct rather than a replacement for radiologists. Tong *et al.* found that an AI-optimized workflow reduced assessment time for experienced radiologists without compromising accuracy (sensitivity: 0.91–1.00, specificity 0.94–0.98), while less experienced practitioners benefited more from a full-AI strategy[19]. According to Zhou *et al.*, AI achieved diagnostic accuracy comparable to FNAB combined with BRAF^V600E testing, including in nodules with indeterminate cytology (Bethesda III/IV)[20]. The performance of our cnn3 or DenseNet models may likewise support decision-making regarding fine-needle aspiration biopsy (FNAB), especially among junior clinicians.

An additional focus of our study was the identification of sonographic features most strongly associated with malignancy risk. As shown in Fig. 4, the presence of deep hypoechogenicity, microcalcifications, and ill-defined margins correlated significantly with thyroid carcinoma. These findings are consistent with existing literature and represent high-risk features underpinning automated risk stratification systems in modified TI-RADS frameworks. For example, Wu *et al.* demonstrated that deep learning enhanced the differentiation of intermediate- and high-risk lesions (TR4 and TR5) within the ACR TI-RADS system, achieving an AUC of up to 0.90 with a significant improvement in sensitivity[21]. Similarly, Wang *et al.* reported that a hybrid model combining ResNet50 and XGBoost achieved higher diagnostic accuracy (76.77%) compared to radiologists (68.38%), with microcalcifications identified as the most predictive feature[22].

The adaptability of AI to local ultrasound classification systems is noteworthy. Li *et al.* showed that modified ACR TI-RADS variants (mACR, mAI) improved specificity, accuracy, and AUC, thereby

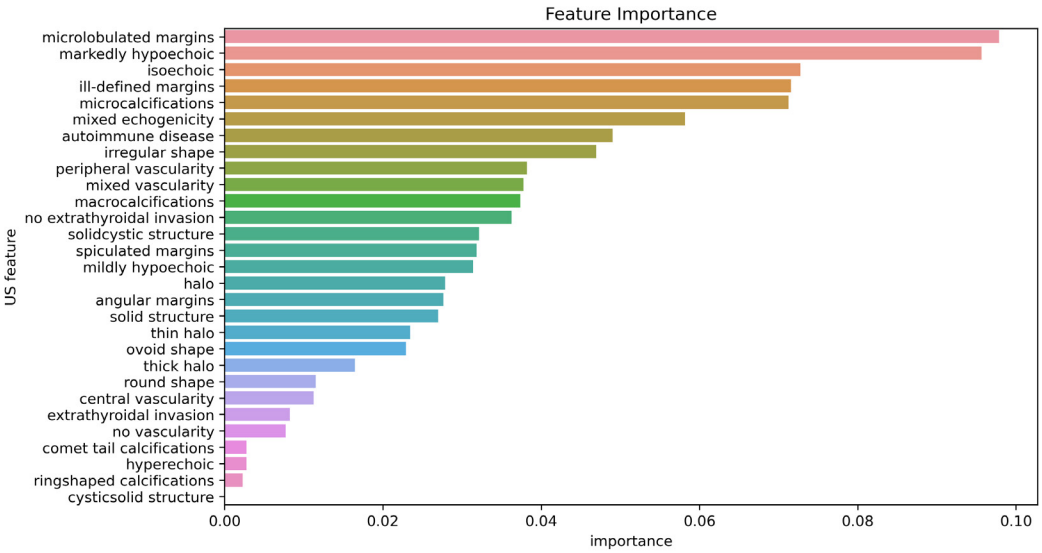

Fig. 4. Importance of US features. Source: own study

reducing unnecessary biopsies without compromising sensitivity[23]. Cui *et al.* confirmed that AI TI-RADS resulted in fewer unnecessary biopsies (41% vs. 47.8%) and fewer missed cancers (22.8% vs. 27.5%; $p < 0.05$) compared to the standard ACR TI-RADS[24].

A more advanced strategy was proposed by Li *et al.*, who introduced a multi-modal approach, combining machine learning models with CNN-based image analysis[25]. Their results showed that machine learning models had an AUC of 0.85, while imaging models based on deep feature extraction reached an AUC of 0.83, and the hybrid model yielded an AUC of 0.87[25]. A feature-based approach was also used by Yao *et al.*, combining CNN models with specific US features[26]. In this case, the features were not indicated by the diagnostician, but provided by other AI models, specialized in identifying these particular features, and the AUC result was 0.88[26]. The advantages of using US features were also emphasized by Gomes Ataide *et al.*, who obtained a sensitivity of 0.92 and a specificity of 0.91, based solely on eleven selected US features[27].

Despite the promising results, this study has several limitations. First, the data are from a single center, which may affect the generalizability of the results. Second, the lack of integration of clinical and molecular data limits the ability to fully stratify patient risk. Additionally, despite the use of cross-validation, the size of the dataset was smaller than in multicenter studies, which may affect the variability of the results. Finally, the models were not tested in a live clinical setting, which should be the subject of further studies.

Conclusions

We extended eight baseline CNN models by integrating five selected ultrasound features which were externally assessed by experienced diagnosticians. This hybrid approach aimed to combine the strengths of deep learning with clinically relevant semantic input. Across all models, the inclusion of ultrasound features led to a consistent improvement in classification performance, yielding an average increase in AUC of approximately 7%.

Conflict of interest

The authors do not report any financial or personal connections with other persons or organizations which might negatively affect the contents of this publication and/or claim authorship rights to this publication.

Author contributions

Original concept of study: AŻ, MR, KDS. Writing of manuscript: AŻ, MR, KDS. Analysis and interpretation of data: AŻ, MR, KDS. Final acceptation of manuscript: AŻ, MR, MD, KDS.
Collection, recording and/or compilation of data: AŻ, MR, KDS. Critical review of manuscript: AŻ, MR, MD, KDS.

References

1. Wojciechowska U, Didkowska J, Barańska K, Miklewska M, Michałek I, Olasek P, Jawołowska A: Nowotwory złośliwe w Polsce w 2022 roku. Krajowy Rejestr Nowotworów. Ministerstwo Zdrowia, Warszawa 2024: 50–51.
2. Alexander EK, Cibas ES: Diagnosis of thyroid nodules. Lancet Diabetes Endocrinol 2022; 10: 533–539. doi: 10.1016/S2213-8587(22)00101-2.
3. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA *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. doi: 10.1016/j.jacr.2017.01.046.
4. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L: European Thyroid Association Guidelines for Ultrasound Malignancy Risk Stratification of Thyroid Nodules in Adults: The EU-TIRADS. Eur Thyroid J 2017; 6: 225–237. doi: 10.1159/000478927.
5. Ha EJ, Na DG, Baek JH: Korean Thyroid Imaging Reporting and Data System: current status, challenges, and future perspectives. Korean J Radiol 2021; 22: 1569–1578. doi: 10.3348/kjr.2021.0106.
6. Friedrich-Rust M, Meyer G, Dauth N, Berner C, Bogdanou D, Herrmann E *et al.*: Interobserver agreement of Thyroid Imaging Reporting and Data System (TI-

RADS) and strain elastography for the assessment of thyroid nodules. PLoS One 2013; 8: e77927. doi: 10.1371/journal.pone.0077927.

7. Yang L, Li C, Chen Z, He S, Wang Z, Liu J: Diagnostic efficiency among Eu-/C-/ACR-TIRADS and S-Detect for thyroid nodules: a systematic review and network meta-analysis. Front Endocrinol (Lausanne) 2023; 14: 1227339. doi: 10.3389/fendo.2023.1227339.
8. Toro-Tobon D, Loor-Torres R, Duran M, Fan JW, Singh Ospina N, Wu Y *et al.*: artificial intelligence in thyroidology: a narrative review of the current applications, associated challenges, and future directions. Thyroid 2023; 33: 903–917. doi: 10.1089/thy.2023.0132.
9. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W: Deep learning to improve breast cancer detection on screening mammography. Sci Rep 2019; 9: 12495. doi: 10.1038/s41598-019-48995-4.
10. Zhou H, Liu B, Liu Y, Huang Q, Yan W: Ultrasonic intelligent diagnosis of papillary thyroid carcinoma based on machine learning. J Healthc Eng 2022; 2022: 6428796. doi: 10.1155/2022/6428796. Retraction in: J Healthc Eng 2023; 2023: 9854690. doi: 10.1155/2023/9854690.
11. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C *et al.*; Collaborators. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer 2019; 113: 47–54. doi: 10.1016/j.ejca.2019.04.001.
12. Todoroki Y, Iwamoto Y, Lin L, Hu H, Chen YW: Automatic Detection of Focal Liver Lesions in Multi-phase CT Images Using A Multi-channel & Multi-scale CNN. Annu Int Conf IEEE Eng Med Biol Soc 2019; 2019: 872–875. doi: 10.1109/EMBC.2019.8857292.
13. Bradshaw TJ, Huemann Z, Hu J, Rahmim A: A guide to cross-validation for artificial intelligence in medical imaging. Radiol Artif Intell 2023; 5: e220232. doi: 10.1148/ryai.220232.
14. Lee JH, Kim YG, Ahn Y, Park S, Kong HJ, Choi JY *et al.*: Investigation of optimal convolutional neural network conditions for thyroid ultrasound image analysis. Sci Rep. 2023;13: 1360. doi: 10.1038/s41598-023-28001-8.
15. Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J *et al.*: Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol 2019; 20: 193–201. doi: 10.1016/S1470-2045(18)30762-9. Erratum in: Lancet Oncol 2020; 21: e462. doi: 10.1016/S1470-2045(20)30546-5.
16. Lai M, Feng B, Yao J, Wang Y, Pan Q, Chen Y *et al.*: Value of artificial intelligence in improving the accuracy of diagnosing TI-RADS category 4 nodules. Ultrasound Med Biol 2023; 49: 2413–2421. doi: 10.1016/j.ultrasmedbio.2023.08.008.
17. Peng S, Liu Y, Lv W, Liu L, Zhou Q, Yang H *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. doi: 10.1016/S2589-7500(21)00041-8. Erratum in: Lancet Digit Health 2021; 3: e413. doi: 10.1016/S2589-7500(21)00110-2.
18. Namsena P, Songsaeng D, Keatmanee C, Klabwong S, Kunapinun A, Soodchuen S *et al.*: Diagnostic performance of artificial intelligence in interpreting thyroid nodules on ultrasound images: a multicenter retrospective study. Quant Imaging Med Surg 2024; 14: 3676–3694. doi: 10.21037/qims-23-1650.
19. Tong WJ, Wu SH, Cheng MQ, Huang H, Liang JY, Li CQ *et al.*: Integration of artificial intelligence decision aids to reduce workload and enhance efficiency in thyroid nodule management. JAMA Netw Open 2023; 6: e2313674. doi: 10.1001/jamanetworkopen.2023.13674.
20. Zhou T, Xu L, Shi J, Zhang Y, Lin X, Wang Y *et al.*: US of thyroid nodules: can AI-assisted diagnostic system compete with fine needle aspiration? Eur Radiol 2024; 34: 1324–1333. doi: 10.1007/s00330-023-10132-1.
21. Wu GG, Lv WZ, Yin R, Xu JW, Yan YJ, Chen RX *et al.*: Deep learning based on ACR TI-RADS can improve the differential diagnosis of thyroid nodules. Front Oncol 2021; 11: 575166. doi: 10.3389/fonc.2021.575166.
22. Wang J, Jiang J, Zhang D, Zhang YZ, Guo L, Jiang Y *et al.*: An integrated AI model to improve diagnostic accuracy of ultrasound and output known risk features in suspicious thyroid nodules. Eur Radiol 2022; 32: 2120–2129. doi: 10.1007/s00330-021-08298-7.
23. Li X, Peng C, Liu Y, Hu Y, Yang L, Yu Y *et al.*: Modified American College of Radiology Thyroid Imaging Reporting and Data System and Modified Artificial Intelligence Thyroid Imaging Reporting and Data System for Thyroid Nodules: a multicenter retrospective study. Thyroid 2024; 34: 88–100. doi: 10.1089/thy.2023.0429.
24. Cui Y, Fu C, Si C, Li J, Kang Y, Huang Y, Cui K: Analysis and comparison of the malignant thyroid nodules not recommended for biopsy in ACR TIRADS and AI TIRADS with a large sample of surgical series. J Ultrasound Med 2023; 42: 1225–1233. doi: 10.1002/jum.16132.
25. Li J, Guo Q, Tan X: Multi-modal feature integration for thyroid nodule prediction: Combining clinical data with ultrasound-based deep features. J Radiat Res Appl Sci 2025; 18: 101217. doi: 10.1016/j.jrras.2024.101217.
26. Yao S, Shen P, Dai T, Dai F, Wang Y, Zhang W, Lu H: Human understandable thyroid ultrasound imaging AI report system – A bridge between AI and clinicians. iScience. 2023; 26: 106530. doi: 10.1016/j.isci.2023.106530
27. Gomes Ataide EJ, Ponugoti N, Illanes A, Schenke S, Kreissl M, Friebe M: Thyroid nodule classification for physician decision support using machine learning-evaluated geometric and morphological features. Sensors (Basel) 2020; 20: 6110. doi: 10.3390/s20216110.