ULTRA SONO GRAPHY

Application of machine learning to ultrasound images to differentiate follicular neoplasms of the thyroid gland

Ilah Shin[1], Young Jae Kim[2], Kyunghwa Han[1], Eunjung Lee[3], Hye Jung Kim[4], Jung Hee Shin[5], Hee Jung Moon[1], Ji Hyun Youk[6], Kwang Gi Kim[2], Jin Young Kwak[1]

[1]Department of Radiology, Severance Hospital, Research Institute of Radiological Science, Yonsei University College of Medicine, Seoul; [2]Department of Biomedical Engineering, Gachon University College of Medicine, Incheon; [3]Department of Computational Science and Engineering, Yonsei University, Seoul; [4]Department of Radiology, Kyungpook National University Chilgok Hospital, School of Medicine, Kyungpook National University, Daegu; [5]Department of Radiology and Center for Imaging Science, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul; [6]Department of Radiology, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Korea

Purpose: This study was conducted to evaluate the diagnostic performance of machine learning in differentiating follicular adenoma from carcinoma using preoperative ultrasonography (US).

Methods: In this retrospective study, preoperative US images of 348 nodules from 340 patients were collected from two tertiary referral hospitals. Two experienced radiologists independently reviewed each image and categorized the nodules according to the 2015 American Thyroid Association guideline. Categorization of a nodule as highly suspicious was considered a positive diagnosis for malignancy. The nodules were manually segmented, and 96 radiomic features were extracted from each region of interest. Ten significant features were selected and used as final input variables in our in-house developed classifier models based on an artificial neural network (ANN) and support vector machine (SVM). The diagnostic performance of radiologists and both classifier models was calculated and compared.

Results: In total, 252 nodules from 245 patients were confirmed as follicular adenoma and 96 nodules from 95 patients were diagnosed as follicular carcinoma. As measures of diagnostic performance, the average sensitivity, specificity, and accuracy of the two experienced radiologists in discriminating follicular adenoma from carcinoma on preoperative US images were 24.0%, 84.0%, and 64.8%, respectively. The sensitivity, specificity, and accuracy of the ANN and SVM-based models were 32.3%, 90.1%, and 74.1% and 41.7%, 79.4%, and 69.0%, respectively. The kappa value of the two radiologists was 0.076, corresponding to slight agreement.

Conclusion: Machine learning-based classifier models may aid in discriminating follicular adenoma from carcinoma using preoperative US.

Keywords: Follicular neoplasm; Ultrasonography; Machine learning; Artificial neural network; Support vector machine

ORIGINAL ARTICLE

https://doi.org/10.14366/usg.19069
pISSN: 2288-5919 • eISSN: 2288-5943
Ultrasonography 2020;39:257-265

Received: November 5, 2019
Revised: February 25, 2020
Accepted: February 29, 2020

Correspondence to:
Jin Young Kwak, MD, PhD, Department of Radiology, Severance Hospital, Research Institute of Radiological Science, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea

Tel. +82-2-2228-7400
Fax. +82-2-2227-8337
E-mail: docjin@yuhs.ac

Kwang Gi Kim, PhD, Department of Biomedical Engineering, Gachon University College of Medicine, 21 Namdong-daero 774 beon-gil, Namdong-gu, Incheon 21565, Korea

Tel. +82-32-820-4036
Fax. +82-32-460-2361
E-mail: kimkg@gachon.ac.kr

How to cite this article:
Shin I, Kim YJ, Han K, Lee E, Kim HJ, Shin JH, et al. Application of machine learning to ultrasound images to differentiate follicular neoplasms of the thyroid gland. Ultrasonography. 2020 Jul;39(3):257-265.

Introduction

Follicular neoplasms of the thyroid gland are divided into benign follicular adenoma and malignant follicular carcinoma. The differential diagnosis between these two entities is made by identifying the presence of capsular, vascular or extrathyroidal tissue invasion, and nodal or distant metastasis [1]. Thus, in cases without overt extrathyroidal tissue invasion or nodal/distant metastasis on the preoperative examination, the differential diagnosis is made by a pathologic examination after surgical excision [2]. The prevalence of follicular adenoma in patients initially diagnosed with follicular neoplasm is roughly 80%, meaning that a majority of patients undergo diagnostic thyroid lobectomy despite having a benign condition [3,4]. Therefore, there is an evident need to distinguish these two entities preoperatively to avoid this overtreatment of patients with benign disease.

Certain grayscale ultrasonography (US) features, including those that have previously been proposed as malignant US features (hypoechogenicity, noncircumscribed margins, and the presence of calcifications), have shown significant associations with follicular carcinoma compared to follicular adenoma [5–7]. Absence of internal cystic changes, hypoechogenicity, lack of a perilesional halo on US, and larger size have also been shown to be associated with follicular carcinoma as distinct from follicular adenoma [8,9]. However, a majority of follicular carcinomas fail to show the proposed imaging findings, which have low positive predictive values (ranging from 55.6% to 61.2%) for differentiating benign follicular adenoma from malignant follicular carcinoma [5,8].

Machine learning is a new field in medical imaging that has emerged and become the topic of intense interest based on the belief that medical images contain crucial information-some of which seems to be beyond the perception of the human eye-about the underlying physiology of tumors [10,11]. Thus, machine learning is expected to play an important role in precision oncology as a robust, non-invasive method to reveal the characteristics of individual tumors based on medical images. Machine learning is a collective term comprising multiple computational methods and models that extract meaningful features from medical images, and it has been increasingly applied in the field of radiology [12,13]. Several classifier models using various machine learning algorithms have also been applied to thyroid US imaging [14–17]. However, previous studies using classic radiologic lexicons as input variables for several classifier models showed contradictory diagnostic performance in differentiating benign and malignant thyroid nodules compared to experienced radiologists [14,15].

To date, no study has applied machine learning to differentiate follicular adenoma and follicular carcinoma based on their preoperative US findings, a task that is currently considered to be a diagnostic challenge [4,18,19]. In this study, we investigated the utility of machine learning, using support vector machine (SVM) and artificial neural network (ANN)-based models to differentiate follicular adenoma from follicular carcinoma on preoperative US images.

Materials and Methods

Subjects

Patients from two tertiary referral hospitals (Severance Hospital and Samsung Medical Center) of South Korea were included in our study. The Institutional Review Board approved this retrospective study and waived the requirement for informed consent for both study populations.

We reviewed the data of consecutively enrolled patients from January 2012 to December 2015 who were surgically confirmed as having follicular adenoma or carcinoma equal to or larger than 1 cm in diameter. There were 104 nodules in 104 patients from Severance Hospital and 244 nodules in 236 patients from Samsung Medical Center. A total of 348 nodules from 340 patients (261 women and 79 men) were included from two institutes, of which 252 nodules from 245 patients were confirmed as follicular adenoma and 96 nodules from 95 patients were diagnosed as follicular carcinoma (Fig. 1). Among the nodules diagnosed as follicular carcinoma, eight (8.3%) were diagnosed as widely invasive and 87 (91.6%) as minimally invasive.

Visual Analysis of the Nodules by Radiologists

Preoperative US images were retrospectively reviewed by two experienced radiologists, both with 15 years (H.J.M. and H.J.K.) of experience in thyroid imaging and both of whom were blind to the patients' clinical information and histopathologic results. For each nodule, the radiologists selected one category for each of the following five US features: composition (solid, predominantly solid, predominantly cystic, and spongiform), echogenicity (hyperechoic, isoechoic, hypoechoic, and markedly hypoechoic), margin (well-circumscribed, microlobulated, and irregular), calcification (microcalcification, macrocalcification, egg-shell calcification, and absence of calcification), and shape (parallel and non-parallel). The 2015 American Thyroid Association (ATA) guideline was used to stratify each thyroid nodule according to its US pattern as very low suspicion, low suspicion, intermediate suspicion, and high suspicion based on the above features [20]. Categorization of a nodule as highly suspicious was considered to indicate a positive diagnosis for malignancy.

Image Segmentation and Pre-processing

All preoperative US images of the thyroid nodules were collected as grayscale images on the picture archiving and communication system by one of 33 radiologists with 1–22 years of experience in thyroid imaging. The images of the study populations at each institution were exported and viewed in the Paint program in Windows 7 (Microsoft, Redmond, WA, USA). All images for each nodule were reviewed by a radiologist with 20 years of clinical experience (J.Y.K.) and a representative image was retrospectively selected for each nodule. A region of interest (ROI) was manually drawn on the representative image of each nodule by an experienced radiologist (J.Y.K.). The overall workflow is summarized in Fig. 2.

Feature Extraction and Selection

In this study, we used in-house developed software for computerized feature analysis of the US images and machine learning. This program was developed using C/C++ in Microsoft Visual Studio (2010 version, Microsoft).

A total of 96 image features were extracted from the ROIs of thyroid nodules by 2-dimensional image analysis software (ImageJ, National Institutes of Health, Bethesda, MD, USA). Each feature list is summarized and is shown in Supplementary Table 1. The texture features were classified into the following four subgroups according to the extraction method and their intrinsic characteristics: gray-level co-occurrence matrix (GLCM), gray-level run length matrix (GLRLM), Gabor, and Haar wavelet [21,22]. In the GLCM and GLRLM methods, a matrix was created for each of the four directions, while the Gabor method considered four directions and three scales. Two-

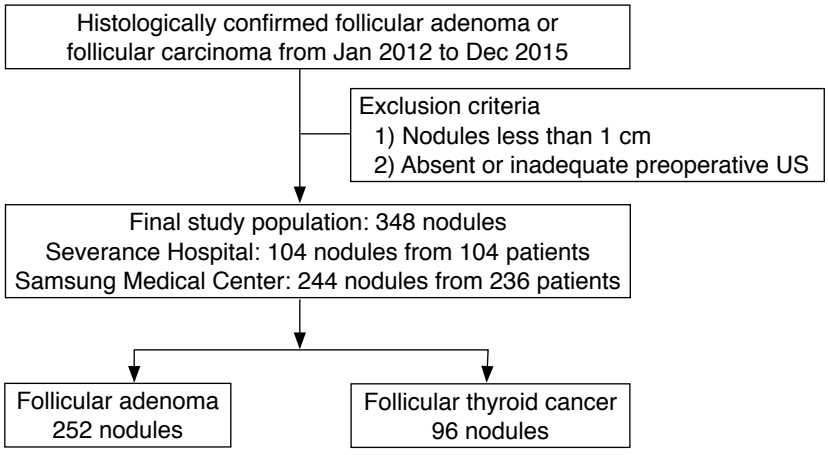


Fig. 1. Inclusion diagram of patient population. US, ultrasonography.

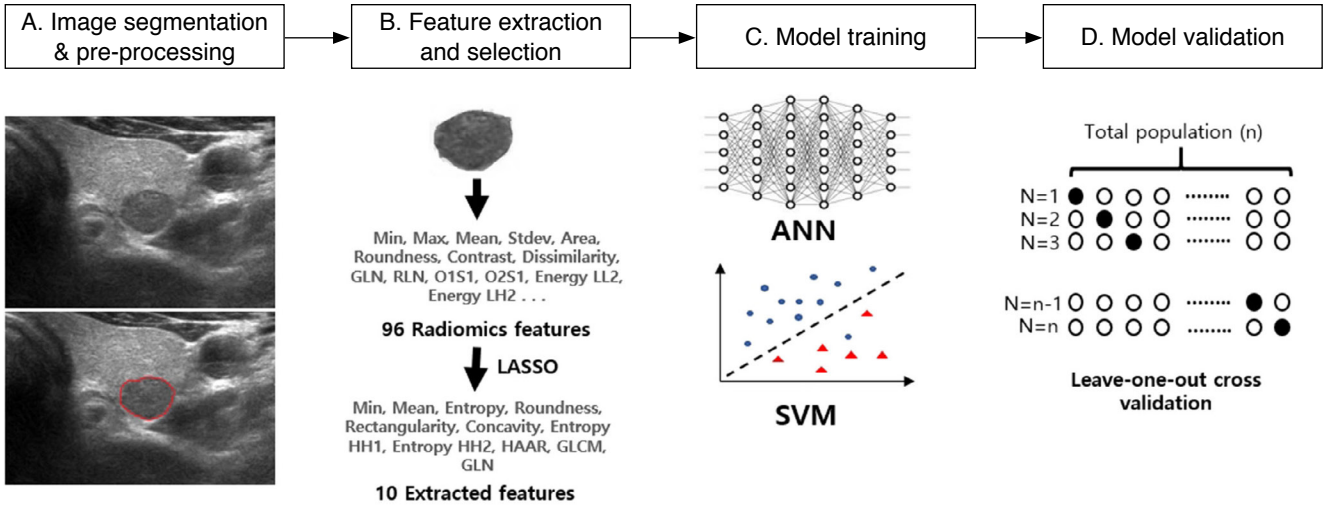


Fig. 2. Diagram of overall workflow of model training and validation. GLN, gray level nonuniformity; RLN, run length nonuniformity; LASSO, least absolute shrinkage and selection operator; GLCM, gray-level co-occurrence matrix; ANN, artificial neural network; SVM, support vector machine.

Table 1. Demographic data of patients

	Total	Follicular adenoma	Follicular carcinoma	P-value
No. of nodules (%)	348	252 (72.4)	96 (27.6)	
Age (yr)	47.2±14.4	47.4±14.0	46.7±15.2	0.671
Size of nodule (mm)[a]	31.0±1.7	29.0 (17.0–40.0)	29.5 (18.0–45.0)	0.261
Male sex	79	54 (21.4)	25 (26.0)	0.359

Values are presented as mean±standard deviation or number (%) unless otherwise indicated.
[a]Median values are shown, with interquartile values of 25% and 75% in parentheses.

level wavelet transformation was done in the Haar wavelet analysis. A total of seven sub-band decompositions were performed, and energy and entropy were extracted for each band. Each extracted feature was represented in different ranges; to solve this problem, the feature values were normalized to values between 0 and 1 by the min-max method.

A statistical selection process was performed to identify significant candidates among the extracted features. The least absolute shrinkage and selection operator (LASSO) method was used to select features [23]. Ten features were finally selected for use as input variables of the classifier models. The selected features were as follows: min, mean, entropy, and 0° contrast from the GLCM features; 0° gray level nonuniformity from the GLRLM features; roundness, rectangularity, and concavity from the morphology features; and entropy (HH2) and entropy (HH1) from the Haar features. The extracted features from the preoperative US images of 252 surgically proven follicular adenoma nodules and 96 follicular carcinoma nodules were implemented in our in-house developed SVM and ANN classifiers.

Classification Model and Validation

The classifier models were built using in-house developed software. We applied two classification algorithms-an ANN and SVM-to classify our data. The SVM calculated the optimal hyperplane using a linear classification model and classified it into two classes [24]. The ANN model had a feed-forward architecture and was trained by using the back-propagation algorithm with the hyperbolic tangent activation function [25]. The ANN model consisted of an input layer of 10 neurons, a hidden layer of 12 neurons, and an output layer of two neurons. Since the training data size was small, model validation was done by using the leave-one-out cross-validation method.

Statistical Analysis

Demographic data on patients' age and sex were collected for each subgroup of follicular adenoma and follicular carcinoma. The independent two-sample t test and chi-square test were used to compare these two variables, respectively. The Mann-Whitney U test was done to compare the mean nodule diameter between the follicular adenoma and carcinoma subgroups.

Sensitivity, specificity, and accuracy were calculated to quantify the performance of radiologists referring to the 2015 ATA guideline and each classifier model for discriminating between follicular adenoma and carcinoma. Sensitivity, specificity, and accuracy were compared using logistic regression with generalized estimating equations. Area under the receiver operating characteristic curve (AUC) values were measured for both radiologists and both classifier models. To consider data clustering caused by patients having multiple thyroid nodules, AUC values were compared and 95% confidence intervals (95% CIs) were calculated using the Obuchowski method [26]. Additionally, a cross-validated AUC was derived during model construction. The radiologists' average values of sensitivity, specificity, and accuracy were also derived and compared with the corresponding values of the classifier models using logistic regression with generalized estimating equations.

Cohen kappa coefficients were derived to compare the interobserver agreement of the visual analysis of the two radiologists. The bootstrap method with 1,000 resamples was used to derive the 95% CI. Kappa values of 0–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80, and 0.81–1.00 were considered to indicate slight agreement, fair agreement, moderate agreement, good agreement, and perfect agreement, respectively [27]. Positive and negative percent agreement were also calculated, considering the unbalanced and asymmetric nature of our study population

The statistical analysis was performed using R version 3.4.2 (R Foundation for Statistical Computing, Vienna, Austria). P-values less than 0.05 were considered to indicate statistical significance.

Results

Subjects

The mean age of patients was 47.2 years (range, 11 to 85 years; standard deviation, 14.4 years) and the mean size of the nodules was 3.1 cm (range, 1.0 to 15.0 cm; standard deviation, 1.7 cm). The demographic data, including patients' age, nodule size, and the sex ratio, showed no significant differences between follicular adenoma and carcinoma (Table 1). Thirteen patients had two nodules: eight patients had two follicular adenomas, four patients had both a follicular adenoma and a follicular carcinoma, and one patient had two follicular carcinomas.

Diagnostic Performance of the Radiologists and Classifier Models for Nodule Classification

The average diagnostic performance values of the two radiologists referring to the 2015 ATA guideline were calculated (Table 2). The sensitivity, specificity, and accuracy of radiologist 1 were 3.1%, 94.8%, and 69.5%, respectively. The results for radiologist 2 were 44.8%, 65.9%, and 60.1%, respectively. All values showed significant differences between the radiologists (sensitivity, P<0.001; specificity, P<0.001; and accuracy, P=0.003). The reader-averaged sensitivity, specificity, and accuracy were 24.0%, 80.4%, and 64.8%, respectively. The AUC values of radiologist 1 and 2 were 0.490 (95% CI, 0.468 to 0.512) and 0.553 (95% CI, 0.495 to 0.612), respectively, and were significantly different from each other (P=0.038).

The diagnostic performance of both classifier models was derived and compared with the radiologists' average values (Table 2). The ANN classifier model showed an accuracy of 74.1%, with a sensitivity of 32.3% and a specificity of 90.1%. Similarly, the SVM classifier model showed an accuracy of 69.0%, a sensitivity of 41.7%, and a specificity of 79.4%. Both classifier models showed higher accuracy than the radiologists' average values, with statistical significance for the ANN model (P<0.001). The cross-validated AUC values of the ANN and SVM models were 0.646 (95% CI, 0.544 to 0.653) and 0.599 (95% CI, 0.597 to 0.707), respectively. The AUC values of the ANN and SVM classifier models were 0.612 (95% CI, 0.561 to 0.662) and 0.605 (95% CI, 0.550 to 0.661), respectively. Since a reader-averaged value for AUC cannot be derived, the AUC for each radiologist was compared with the values of each classifier models. The AUC of the ANN classifier model was higher than those of radiologist 1 and radiologist 2 (P<0.001 and P=0.085, respectively). The AUC of the SVM classifier model was also higher than those of radiologist 1 and radiologist 2 (P<0.001 and P=0.146, respectively). Example images of cases with discrepancies are shown in Figs. 3 and 4.

Interobserver Variability

The kappa value was 0.076 (95% CI, 0.017 to 0.139), showing slight agreement between the two radiologists. The overall percent agreement of the two radiologists referring to the ATA guideline was 64.7% (225 of 348). The underlying positive percent agreement was 3.2% (11 of 348) and the negative percent agreement was 61.5% (214 of 348).

Discussion

Preoperative US and fine-needle aspiration cytology have been used to differentiate benign and malignant thyroid nodules, with good diagnostic performance in preoperatively distinguishing papillary thyroid cancer [28,29]. However, these methods play a limited role in discriminating follicular adenoma and carcinoma of the thyroid gland. Certain US features, such as solid appearance, hypoechogenicity, the presence of calcifications, absence of a US halo, and a noncircumscribed margin, have been associated with

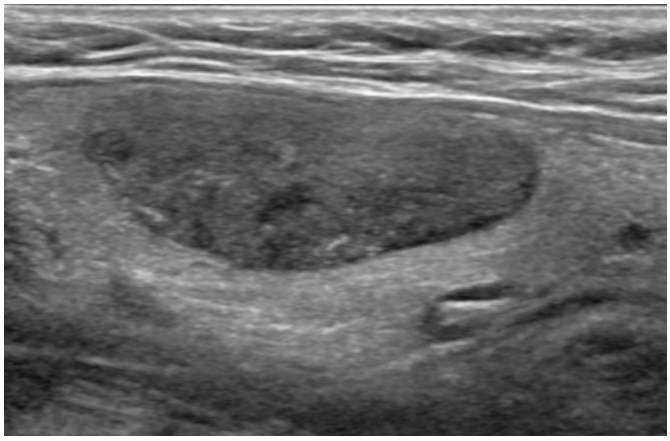
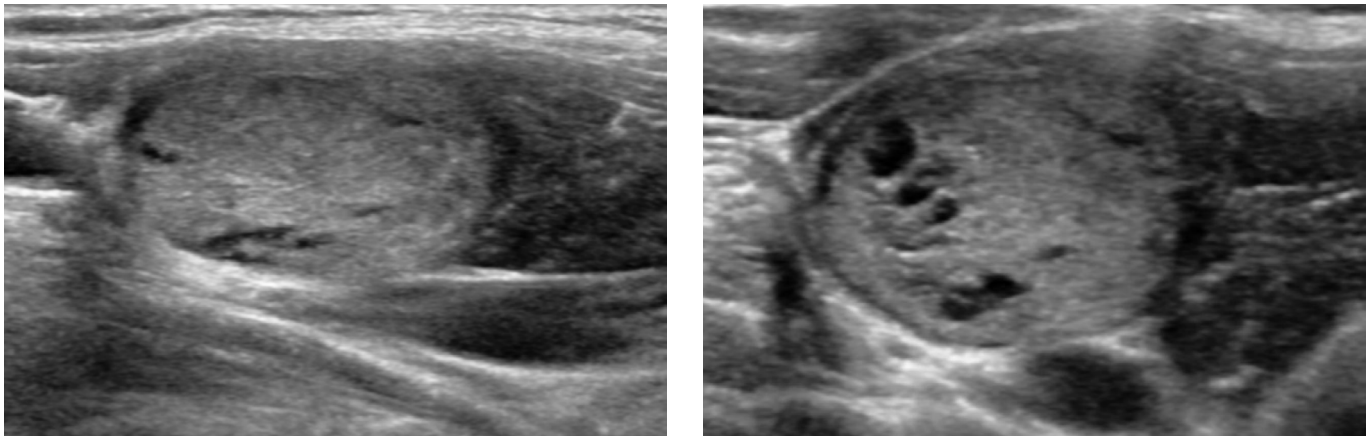


Fig. 3. **Ultrasonography of a 47-year-old woman with pathologically proven follicular adenoma.** The lesion was correctly categorized as benign by both classifier models but was interpreted as malignant by both radiologists.

Table 2. **Diagnostic performance of two radiologists, the radiologists' average values, and the classifier models for differentiating follicular thyroid neoplasms on US**

	Radiologist 1	Radiologist 2	Radiologists' average value	SVM	ANN	P-value[a]	P-value[b]
Sensitivity (%)	3.1	44.8	24.0	41.7	32.3	0.002	0.103
Specificity (%)	94.8	65.9	80.4	79.4	90.1	0.744	<0.001
Accuracy (%)	69.5	60.1	64.8	69.0	74.1	0.137	<0.001
AUC	0.490	0.553	–	0.605	0.612	–	–

US, ultrasonography; SVM, support vector machine; ANN, artificial neural network, AUC, area under the curve.
[a]P-value obtained by comparing the corresponding values with the SVM model and the radiologists' average values. [b]P-value obtained by comparing the corresponding values with the ANN model and the radiologists' average values.



A B

Fig. 4. Ultrasonography of a 33-year-old woman with pathologically proven follicular carcinoma minimally invasive.
On longitudinal (A) and transverse (B) images, the lesion was correctly categorized as malignant by both classifier models but was interpreted as benign by both radiologists.

follicular carcinoma compared to follicular adenoma. However, these features show limited diagnostic performance, with either high sensitivity but low specificity (solid appearance, sensitivity ranging from 68.0% to 90.0% and specificity ranging from 13.7% to 30.8%) or high specificity but low sensitivity (the presence of calcifications and a noncircumscribed margin, sensitivity ranging from 10.2% to 32.6% and specificity ranging from 85.1% to 90.9%) in discriminating follicular carcinoma from adenoma [5,7–9]. Similarly, although certain subtypes of adenoma, such as macrofollicular-type adenoma, may be distinguished from follicular carcinoma by fine-needle aspiration cytology, other subtypes of adenoma such as Hürthle cell adenoma are known to be indistinguishable from follicular carcinoma, meaning that a significant gray zone exists [18,30–32]. Core needle biopsy and intraoperative frozen sections discriminate follicular adenoma from carcinoma with slightly high specificity and low sensitivity, and the frequent indeterminate results of these methods hinder their practical use as independent tools [33,34].

In our study, we developed radiomics-based classifier models to differentiate follicular adenoma and carcinoma based on preoperative US images. The diagnostic performance of our models was evaluated and compared with that of experienced radiologists, who categorized each nodule according to the 2015 ATA guideline. Nodules classified as highly suspicious were considered to have received a positive diagnosis for malignancy. In this setting, our radiomics-based classifier models showed higher overall accuracy than the experienced radiologists (radiologist average, 64.8%; SVM, 69.0%; ANN, 74.0%). Additionally, our radiomics-based classifier models showed relatively high specificity (79.4% and 90.1% for SVM and ANN, respectively) in discriminating follicular carcinoma and adenoma. Although the overall accuracy of our radiomics-based model is limited, it could be used as an adjuvant tool to preoperatively differentiate follicular adenoma and carcinoma. To our knowledge, this is the first study to apply radiomics to preoperative US to predict malignancy in a study population exclusively including follicular neoplasms of the thyroid gland. Several previous studies have applied radiomics to predict the malignancy of thyroid nodules on US, but have done so regardless of histologic subtype [35]. Liang et al. [36] included 137 thyroid nodules (52 benign and 82 malignant nodules) in their training cohort and developed a formula to calculate a radiomics score for each nodule using radiomics features extracted from preoperative US images. Similarly to our study, 1,044 features were initially extracted and then reduced to 19 features using the LASSO regression model. Their radiomics score model had an AUC of 0.921 for predicting malignancy, showing better performance than experienced and junior radiologists referring to the 2017 Thyroid Imaging, Reporting, and Data System scoring criteria [36]. Another study by Yu et al. [17] included 610 thyroid nodules (403 benign and 207 malignant). Texture features were extracted from each nodule and were used to train ANN and SVM-based classifier models to predict malignancy. The ANN and SVM models showed 90.0% and 86.0% accuracy in predicting malignancy, respectively [18]. The radiomics-based models in our study including only follicular neoplasms showed inferior diagnostic performance compared to other radiomics models in the previously mentioned studies, which included all thyroid nodules. However, the experienced radiologists in our study also showed poorer performance in predicting malignancy than the radiologists

in those studies. The reason for this may be differences in the conformity of US findings for predicting follicular carcinoma [37]. A significant gray zone exists in US findings between follicular adenoma and carcinoma, and therefore, there may potentially be a more substantial role for machine learning-based classifier models in discriminating follicular adenoma from carcinoma in larger confirmative studies.

The interobserver variability in discriminating malignant from benign thyroid nodules has overall shown substantial agreement (κ=0.61–0.79) among experienced radiologists [38–40]. To date, no study has evaluated the interobserver variability of US assessment results limited to follicular neoplasms of the thyroid gland on cytology. In our study, the interobserver variability between the two radiologists was poor, with only slight agreement (κ=0.076), even though both radiologists had more than 10 years of experience with thyroid US. Similarly, all performance variables (sensitivity, specificity, and accuracy) for each radiologist showed significantly different values from one another, even though the same guideline was used as a reference point for decision-making. These findings suggest that US-based analyses seeking to discriminate follicular adenoma from carcinoma are much more subjective, with a significant gray zone that yields low reproducibility. Therefore, radiomics-based classifier models using quantitative information from US have the potential to provide more objective results in the preoperative discrimination of follicular adenoma and carcinoma on US.

There are several limitations in our study. First, the study population was small, with a total of 348 nodules consisting of 252 follicular adenomas and 96 follicular carcinomas. Due to this small study population, the leave-one-out cross-validation method was used for model validation, rather than creating a separate validation set. A further assessment with a larger study population should be conducted. Second, demographic information was not applied as input data in our classifier models. Clinical data such as age, sex, and tumor size could be predictors of malignancy in follicular neoplasms of the thyroid [8]. However, in our study, demographic data showed no significant difference between the benign and malignant subgroups, and these variables were therefore not included as input variables in the classifier system. Larger data sets may reveal the potential role of demographic data in diagnosing follicular neoplasms of the thyroid gland. Third, external validation was not done in our study. Due to the low prevalence of follicular neoplasms and the even lower incidence of follicular thyroid carcinoma, it was difficult to prepare a separate group of patients for external validation. Further studies with larger sample sizes are needed for further validation. Fourth, the diagnostic performance and interobserver variability were questionably low, even though both radiologists in this study had more than 10 years of experience in thyroid imaging. This result may have been due to the fact that differentiating follicular adenoma and carcinoma on US is challenging and moreover, the overwhelming majority of the patients in the carcinoma group had minimally invasive tumors (88 of 96, 91.6%), which are especially difficult to distinguish from adenoma. Lastly, our study only utilized one criterion for the US-based diagnosis of malignancy. However, there are currently various guidelines for thyroid nodule characterization, with different thresholds for malignancy in each guideline. Moreover, the majority of guidelines focus on discriminating the more common papillary thyroid cancer, and thus might not be appropriate as a guideline for radiologists in differentiating follicular neoplasms. A further comparison of diagnostic performance with reference to various guidelines should be done in future studies.

Our in-house developed SVM and ANN classifier models, which used texture features as input variables, showed high specificity in differentiating follicular carcinoma from adenoma, with comparable diagnostic performance to that of experienced radiologists referring to the ATA guideline. This is a preliminary study, and with further validation and refinement, classifier models may have the potential to aid attending radiologists in differentiating thyroid follicular neoplasms.

ORCID: Ilah Shin: https://orcid.org/0000-0002-2046-9426; Young Jae Kim: https://orcid.org/0000-0003-0443-0051; Kyunghwa Han: https://orcid.org/0000-0002-5687-7237; Eunjung Lee: https://orcid.org/0000-0001-9989-3555; Hye Jung Kim: https://orcid.org/0000-0002-0263-0941; Jung Hee Shin: https://orcid.org/0000-0001-6435-7357; Hee Jung Moon: https://orcid.org/0000-0002-5643-5885; Ji Hyun Youk: https://orcid.org/0000-0002-7787-780X; Kwang Gi Kim: https://orcid.org/0000-0001-9714-6038; Jin Young Kwak: https://orcid.org/0000-0002-6212-1495

Author Contributions

Conceptualization: Kwak JY, Shin I, Kim YJ, Kim KG. Data acquisition: Kwak JY, Shin JH. Data analysis or interpretation: Kwak JY, Shin I, Kim YJ, Han K, Lee E, Kim HJ, Moon HJ, Youk JH, Kim KG. Drafting of the manuscript: Shin I, Kim YJ. Critical revision of the manuscript: Kwak JY, Shin I, Kim KG. Approval of the final version of the manuscript: all authors.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Supplementary Material

Supplementary Table 1. Ninety-six extracted features from ultrasonography of the thyroid nodules (https://doi.org/10.14366/usg.19069).

References

1. Goldstein RE, Netterville JL, Burkey B, Johnson JE. Implications of follicular neoplasms, atypia, and lesions suspicious for malignancy diagnosed by fine-needle aspiration of thyroid nodules. Ann Surg 2002;235:656-662.
2. St Louis JD, Leight GS, Tyler DS. Follicular neoplasms: the role for observation, fine needle aspiration biopsy, thyroid suppression, and surgery. Semin Surg Oncol 1999;16:5-11.
3. Choi YJ, Yun JS, Kim DH. Clinical and ultrasound features of cytology diagnosed follicular neoplasm. Endocr J 2009;56:383-389.
4. McHenry CR, Phitayakorn R. Follicular adenoma and carcinoma of the thyroid gland. Oncologist 2011;16:585-593.
5. Yoon JH, Kim EK, Youk JH, Moon HJ, Kwak JY. Better understanding in the differentiation of thyroid follicular adenoma, follicular carcinoma, and follicular variant of papillary carcinoma: a retrospective study. Int J Endocrinol 2014;2014:321595.
6. Gulcelik NE, Gulcelik MA, Kuru B. Risk of malignancy in patients with follicular neoplasm: predictive value of clinical and ultrasonographic features. Arch Otolaryngol Head Neck Surg 2008;134:1312-1315.
7. Kuo TC, Wu MH, Chen KY, Hsieh MS, Chen A, Chen CN. Ultrasonographic features for differentiating follicular thyroid carcinoma and follicular adenoma. Asian J Surg 2020;43:339-346.
8. Sillery JC, Reading CC, Charboneau JW, Henrichsen TL, Hay ID, Mandrekar JN. Thyroid follicular carcinoma: sonographic features of 50 cases. AJR Am J Roentgenol 2010;194:44-54.
9. Zhang JZ, Hu B. Sonographic features of thyroid follicular carcinoma in comparison with thyroid follicular adenoma. J Ultrasound Med 2014;33:221-227.
10. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016;278:563-577.
11. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48:441-446.
12. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics 2017;37:505-515.
13. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts H. Machine learning methods for quantitative radiomic biomarkers. Sci Rep 2015;5:13087.
14. Lim KJ, Choi CS, Yoon DY, Chang SK, Kim KK, Han H, et al. Computer-aided diagnosis for the differentiation of malignant from benign thyroid nodules on ultrasonography. Acad Radiol 2008;15:853-858.
15. Wu H, Deng Z, Zhang B, Liu Q, Chen J. Classifier model based on machine learning algorithms: application to differential diagnosis of suspicious thyroid nodules via sonography. AJR Am J Roentgenol 2016;207:859-864.
16. Nam SJ, Yoo J, Lee HS, Kim EK, Moon HJ, Yoon JH, et al. Quantitative evaluation for differentiating malignant and benign thyroid nodules using histogram analysis of grayscale sonograms. J Ultrasound Med 2016;35:775-782.
17. Yu Q, Jiang T, Zhou A, Zhang L, Zhang C, Xu P. Computer-aided diagnosis of malignant or benign thyroid nodes based on ultrasound images. Eur Arch Otorhinolaryngol 2017;274:2891-2897.
18. Baloch ZW, Fleisher S, LiVolsi VA, Gupta PK. Diagnosis of "follicular neoplasm": a gray zone in thyroid fine-needle aspiration cytology. Diagn Cytopathol 2002;26:41-44.
19. Najafian A, Olson MT, Schneider EB, Zeiger MA. Clinical presentation of patients with a thyroid follicular neoplasm: are there preoperative predictors of malignancy? Ann Surg Oncol 2015;22:3007-3013.
20. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, 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.
21. Castellano G, Bonilha L, Li LM, Cendes F. Texture analysis of medical images. Clin Radiol 2004;59:1061-1069.
22. Liu C, Wechsler H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans Image Process 2002;11:467-476.
23. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Methodol 1996;58:267-288.
24. Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett 1999;9:293-300.
25. Orr GB, Muller KR. Neural networks: tricks of the trade. Heidelberg: Springer, 2003.
26. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. Biometrics 1997;53:567-578.
27. Kundel HL, Polansky M. Measurement of observer agreement. Radiology 2003;228:303-308.
28. Moon WJ, Jung SL, Lee JH, Na DG, Baek JH, Lee YH, et al. Benign and malignant thyroid nodules. US differentiation: multicenter retrospective study. Radiology 2008;247:762-770.
29. Bongiovanni M, Spitale A, Faquin WC, Mazzucchelli L, Baloch ZW. The Bethesda System for Reporting Thyroid Cytopathology: a meta-analysis. Acta Cytol 2012;56:333-339.
30. Mazzaferri EL. Management of a solitary thyroid nodule. N Engl J Med 1993;328:553-559.
31. Greaves TS, Olvera M, Florentine BD, Raza AS, Cobb CJ, Tsao-Wei DD, et al. Follicular lesions of thyroid: a 5-year fine-needle aspiration experience. Cancer 2000;90:335-341.
32. Caraway NP, Sneige N, Samaan NA. Diagnostic pitfalls in thyroid fine-needle aspiration: a review of 394 cases. Diagn Cytopathol

1993;9:345-350.
33. Min HS, Kim JH, Ryoo I, Jung SL, Jung CK. The role of core needle biopsy in the preoperative diagnosis of follicular neoplasm of the thyroid. APMIS 2014;122:993-1000.
34. Callcut RA, Selvaggi SM, Mack E, Ozgul O, Warner T, Chen H. The utility of frozen section evaluation for follicular thyroid lesions. Ann Surg Oncol 2004;11:94-98.
35. Sollini M, Cozzi L, Chiti A, Kirienko M. Texture analysis and machine learning to characterize suspected thyroid nodules and differentiated thyroid cancer: where do we stand? Eur J Radiol 2018;99:1-8.
36. Liang J, Huang X, Hu H, Liu Y, Zhou Q, Cao Q, et al. Predicting malignancy in thyroid nodules: radiomics score versus 2017 American College of Radiology Thyroid Imaging, Reporting and Data System. Thyroid 2018;28:1024-1033.
37. Jeh SK, Jung SL, Kim BS, Lee YS. Evaluating the degree of conformity of papillary carcinoma and follicular carcinoma to the reported ultrasonographic findings of malignant thyroid tumor. Korean J Radiol 2007;8:192-197.
38. Choi SH, Kim EK, Kwak JY, Kim MJ, Son EJ. Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. Thyroid 2010;20:167-172.
39. Koh J, Moon HJ, Park JS, Kim SJ, Kim HY, Kim EK, et al. Variability in interpretation of ultrasound elastography and gray-scale ultrasound in assessing thyroid nodules. Ultrasound Med Biol 2016;42:51-59.
40. Grani G, Lamartina L, Cantisani V, Maranghi M, Lucia P, Durante C. Interobserver agreement of various thyroid imaging reporting and data systems. Endocr Connect 2018;7:1-7.