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

Ilah Shin

Department of Medicine

The Graduate School, Yonsei University

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

Ilah Shin

Department of Medicine

The Graduate School, Yonsei University

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

Directed by Professor Jin Young Kwak

Doctoral Dissertation
submitted to the Department of Medicine,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree
of Doctor of Medicine

Ilah Shin

June 2020

This certifies that the Doctoral Dissertation of Ilah Shin is approved.

[Signature]

Thesis Supervisor : Jin Young Kwak

[Signature]

Thesis Committee Member#1 : Ja Seung Koo

[Signature]

Thesis Committee Member#2 : Kyunghwa Han

[Signature]

Thesis Committee Member#3: Kwang Gi Kim

[Signature]

Thesis Committee Member#4: Eunjung Lee

The Graduate School
Yonsei University

June 2020

ACKNOWLEDGEMENTS

I would like to express my gratitude towards my supervisor, professor Jin Young Kwak for all the guidance, constant supervision and the encouragement which helped me in completion of this paper.

I would like to thank Dr. Kyunghwa Han and Dr. Kwang Gi Kim for sharing their knowledge and technical support.

I also wish to express my gratitude to the committee members, professor Ja Seung Goo and Eunjung Lee for sharing their precious time and comment during the period of my paper work.

<TABLE OF CONTENTS>

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

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

Ilah Shin

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Jin Young Kwak)

Purpose : To evaluate the diagnostic performance of machine learning in differentiating follicular adenoma from carcinoma using preoperative ultrasound (US) images.

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 nodule according to the 2015 American Thyroid Association (ATA) guideline. The nodules were manually segmentated and 96 radiomic features were extracted from each region of interest. Ten significant features were selected and used as final input variables to our in-house developed classifier models based on artificial neural network (ANN) and support vector machine (SVM). The diagnostic performance of radiologists and both classifier models were calculated and compared.

Results : Two hundred and fifty-two nodules from 245 patients were confirmed as follicular adenoma and ninety-six nodules from 95 patients were diagnosed as follicular carcinoma. The averaged diagnostic performance of sensitivity, specificity and accuracy of two experienced radiologists in discriminating follicular and adenoma and carcinoma on preoperative US images were 24.0%, 84.0% and 64.8%, respectively. The sensitivity, specificity and accuracy of

ANN and SVM based models were 32.3%, 90.1%, 74.1% and 41.7%, 79.4%, 69.0%, respectively. The kappa value of the two radiologists was 0.076 and showed slight agreement.

Conclusion : Machine learning based classifier models showed better diagnostic performance compared with experienced radiologists in discriminating follicular adenoma and carcinoma using preoperative US images.

Keywords : radiomics, support vector machine, artificial neural network, follicular adenoma, follicular thyroid carcinoma

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

Ilah Shin

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Jin Young Kwak)

I. INTRODUCTION

Follicular neoplasm of the thyroid gland consists of benign follicular adenoma and malignant follicular carcinoma. The differential diagnosis among these two entities is made by identifying presence of capsular, vascular or extrathyroidal tissue invasion and nodal or distant metastasis [1]. And thus, in cases without overt extrathyroidal tissue invasion or nodal/distant metastasis on preoperative examinations, the differential diagnosis is made by pathologic examination after surgical excision [2]. The prevalence of follicular adenoma in patients initially diagnosed as follicular neoplasm is about 80%, meaning a majority of the patients undergo diagnostic thyroid lobectomy due to benign disease [3,4]. So, it is evident that the need to preoperatively distinguish these two entities is crucial to avoid this overtreatment of patients with benign disease.

Gray-scale ultrasound (US) features such as previously proposed malignant US features (marked hypoechogenicity, non-circumscribed margins, microscopic or mixed calcifications, and taller than wide configuration) have been significantly associated with follicular carcinoma compared to follicular adenoma [5,6]. Absence of internal cystic changes, hypoechogenicity, lack of US perilesional halo and larger size have also been shown to be associated with follicular carcinoma compared to follicular adenoma [7]. However, a majority of

follicular carcinoma fail to show the proposed imaging findings, showing low positive predictive values ranging from 55.6% to 61.2% for these imaging findings in differentiating benign follicular adenoma and malignant follicular carcinoma [5-7].

Machine learning is a new field in medical imaging that has emerged in fame through the belief that medical images contain crucial information_some seem to be beyond the perception of human eye_of the underlying tumor physiology [8,9]. And thus, it is expected to have an important role in precision oncology as a robust non-invasive method to reveal individual tumor characteristics through medical images. Machine learning is a collective term comprising group of computational methods/models that extracts meaningful features from medical images, and thus it has been increasingly applied in the field of radiology [10,11]. Several classifier models from various machine learning algorithms had also been applied in thyroid US imaging [12-15]. Some previous studies using classic radiologic lexicons as input variables to several classifier models showed contradictory diagnostic performance in differentiating benign and malignant thyroid nodules compared to the experienced radiologists in each study [12,13].

To date, no study has applied machine learning to differentiate follicular adenoma and follicular carcinoma based on their preoperative US findings. It is currently thought to be challenging to pre-operatively differentiate follicular adenoma and follicular carcinoma [4,16,17]. In this study, we investigated the utility of machine learning – specifically using support vector machine (SVM) and artificial neural network (ANN) based models in differentiating follicular adenoma and follicular carcinoma in preoperative US images.

II. MATERIALS AND METHODS

1. Subjects

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

From January 2012 to December 2015, we reviewed data of consecutively enrolled patients whom were confirmed as follicular adenoma or carcinoma through surgery which were equal to or larger than 1cm in diameter. There were 104 nodules in 104 patients from Severance Hospital and 244 nodules in 236 patients from Samsung Medical Center. Finally, a total of 348 thyroid nodules from 340 patients were collected from the two institutes and included in this study (Figure 1).

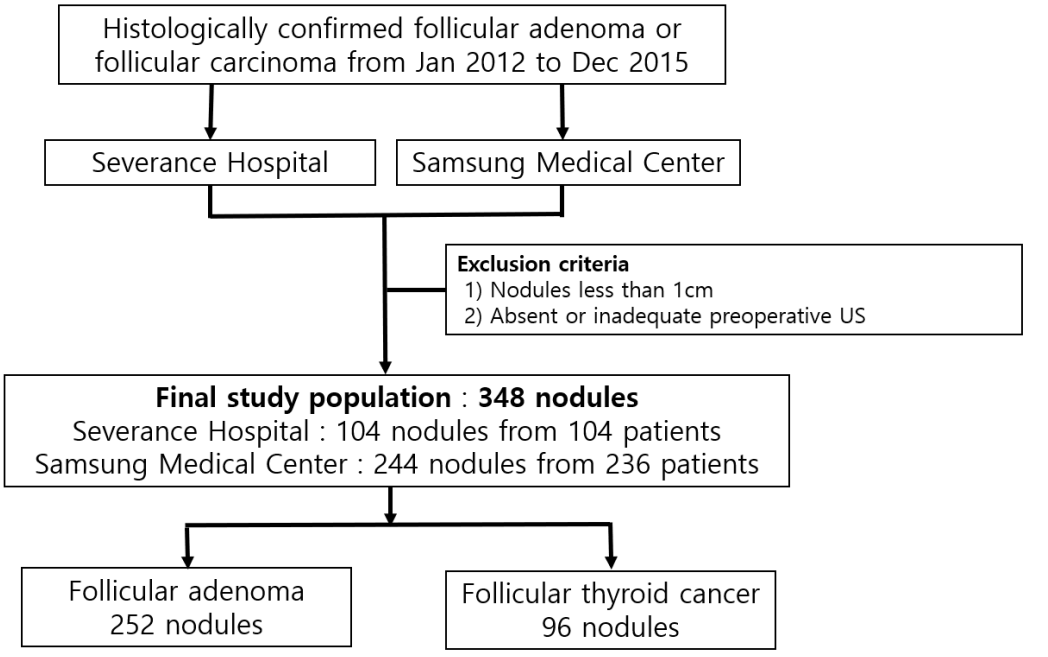


Figure 1. Inclusion diagram of patient population

2. Visual analysis of the nodules by radiologists

Preoperative US images were retrospectively reviewed by two experienced radiologists, both with 15 years (XXX, XXX) of experience in thyroid imaging, both of whom were blind to clinical information and histopathologic results. Five US features of composition, echogenicity, margin, calcification and shape of each nodule were recorded by each radiologist. Each radiologist was asked to select one of the categories for each feature. Categories for each feature were as follows; 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), shape (parallel and non-parallel). The 2015 American Thyroid Association (ATA) guideline was used to stratify each thyroid nodule as either very low suspicious, low suspicious, intermediate suspicious and high suspicious US pattern according to the above features [18]. Nodules categorized in the 'highly suspicious' category were considered as a positive diagnosis for malignancy.

3. Image segmentation and pre-processing

The overall workflow diagram was summarized in Figure 2. All preoperative US images of the thyroid nodules were collected as grayscale images on the picture archiving and communication system (PACS) by one of thirty-three radiologists with 1-22 years of experience in thyroid imaging. Images of the study populations of each institution were both exported and brought up in the Paint program of Windows 7 (Microsoft, Redmond, WA). All

images for each nodule were reviewed by an experienced radiologist (J.Y.K) and the representative image was retrospectively selected for each nodule. Region of interest (ROI) was drawn on the representative image of each nodule manually by the experienced radiologist (J.Y.K).

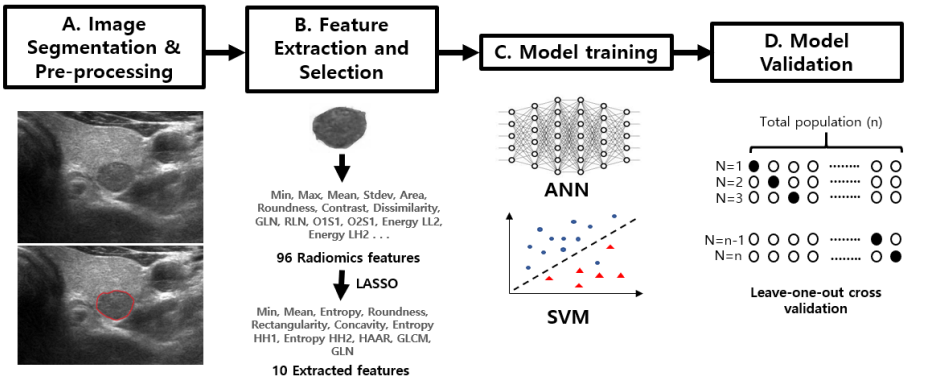


Figure 2. Diagram of overall workflow of model training and validation
LASSO Least absolute shrinkage and selection operator; ANN Artificial neural network; SVM Support vector machine

3. Feature extraction and selection

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

A total of 96 image features were extracted from region of ROIs of thyroid nodules by 2D image analysis software (ImageJ, Bethesda, Maryland, USA). Each feature list is summarized and is shown in S1 Table. The texture features were classified into four subgroups according to the extraction method and intrinsic characteristics; Gray-level co-occurrence matrix (GLCM), Gray-level run length matrix (GLRLM), Gabor, and Haar wavelet [19,20]. GLCM

and GLRLM methods created a matrix for each of the four directions and Gabor method considered four directions and three scales. Two-level wavelet transformation was done in the Haar wavelet analysis. A total of 7 sub-band decompositions were performed, and energy and entropy were extracted for each band. Each extracted feature was represented in different ranges and thus to solve this problem, the feature values are normalized to values between 0 and 1 by MIN-MAX method.

A statistical selection process was performed to select significant candidates among the extracted features. The Least Absolute Shrinkage and Selection Operator (LASSO) method was used to select features [21]. Ten features were finally selected for use as input variables of the classifier models. The selected features included Min, Mean, Entropy, 0 degree contrast from GLCM feature, 0 degree GLN from GLRLM features, roundness, rectangularity and concavity from the morphology features, Entropy (HH2) and Entropy (HH1) from HAAR features. The extracted features from the preoperative US images of surgically proven two hundred fifty-two follicular adenoma nodules and ninety-six follicular carcinoma nodules were implemented into our in-house developed SVM and ANN classifiers.

4. Classification model and validation

The classifier models were built using in-house developed software. We applied two classification algorithms – artificial neural network (ANN) and support vector machine (SVM) – to classify our data. The SVM calculated the optimal hyperplane using a linear classification model and classified it into two classes [22]. The ANN model had feed-forward architecture and was trained by using the back-propagation algorithm with the hyperbolic tangent activation function [23]. The ANN model consisted of an input layer of 10 neurons, a hidden layer of 12 neurons, and an output layer of 2 neurons. Since the training data

size was small, model validation was done by using the leave-one-out cross validation method.

5. Statistical analysis

Demographic data of patient age and sex were collected for each subgroup of follicular adenoma and follicular carcinoma. Independent two-sample t-test and Chi-square test was used to compare these two variables, respectively. Mann-Whitney U test was done to compare the mean nodule diameter between follicular adenoma and carcinoma subgroups.

Sensitivity, specificity and accuracy were calculated to quantify the discrimination performance of radiologists referring to the 2015 ATA guideline and each classifier model. Sensitivity, specificity and accuracy were compared using logistic regression with the generalized estimating equation. The area under curve (AUC) values of the receiver operating characteristic (ROC) curves were measured for both radiologists and both classifier models. To consider data clustering caused by patients having multiple thyroid nodules, comparison of AUC values and calculation of 95% confidence interval (95% CI) were done using Obuchowski's method [24]. Also, cross validated AUC was derived during model construction. Radiologist-averaged values of sensitivity, specificity and accuracy were also derived and compared with corresponding values of the classifier models using logistic regression with the generalized estimating equation.

Cohen's kappa value was derived to compare the interobserver agreement of the visual analysis of two radiologists. Bootstrap method with 1,000 resampling was used to derive the 95% confidence interval. Kappa value of 0-0.20, 0.21-0.40, 0.41-0.60, 0.61-0.80, 0.81-1.00 indicated slight agreement, fair agreement, moderate agreement, good agreement and perfect agreement [25]. Overall, 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 package, version 3.4.2 (R Foundation for Statistical Computing, Vienna, Austria). *P*-values less than 0.05 were considered statistically significant.

III. RESULTS

1. Subjects

A total of 348 nodules from 340 patients (261 women and 79 men) were included from two institutes. Two hundred and fifty-two nodules from 245 patients were confirmed as follicular adenoma and ninety-six nodules from 95 patients were diagnosed as follicular carcinoma. Among the nodules diagnosed as follicular carcinoma, eight nodules (8.3%) were diagnosed as widely invasive and 87 nodules (91.6%) as minimally invasive type. The mean age of patients was 47.2 years old (range 11 to 85 years old, standard deviation 14.4 years) and the mean size of the nodules was 3.1cm (range 1.0 to 15.0cm, standard deviation 1.7cm). The demographic data including age of patients, nodule size and proportion of male sex showed no significant difference among follicular adenoma and carcinoma (Table 1). The 13 patients had 2 nodules – 8 patients had two follicular adenomas, 4 patients had both a follicular adenoma and a follicular carcinoma and one patient had two follicular carcinomas.

Table 1. Demographics of patient population

	Total	**Follicular adenoma**	**Follicular carcinoma**	***p*-value**
No. of nodules	348	252 (72.4%)	96 (27.6%)	
Age (year)	47.2 ± 14.4	47.4 ± 14.0	46.7 ± 15.2	0.671

Size of nodule (mm)*	31.0 ± 1.7	29.00 (17.0, 40.0)	29.5 (18.0, 45.0)	0.261
Male sex	79	54 (21.4%)	25 (26.0%)	0.359

* Median value is shown with interquartile values of 25% and 75% in parenthesis

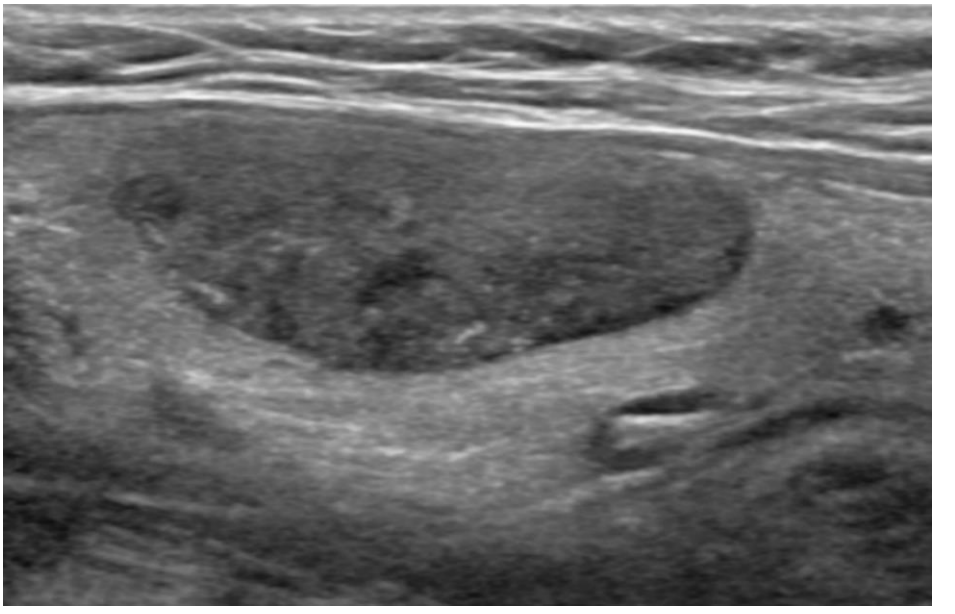
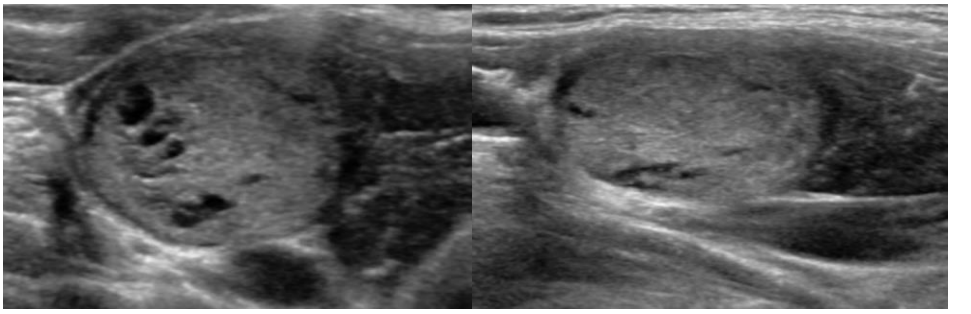


Figure 3. Ultrasound image of a 47-year-old woman with pathologically proven follicular adenoma, of both which were correctly categorized as benign by both classifier models but were interpreted as malignant by both radiologists



(A) **(B)**

Figure 4. Longitudinal (A) and transverse (B) ultrasonography of a 33-year-old female patient with pathologically proven follicular carcinoma minimally invasive, of both which were correctly categorized as malignant by both classifier models but were interpreted as benign by both radiologists.

2. Diagnostic performances of the radiologists and classifier models for nodule classification

The diagnostic performance values of two radiologists referring to the 2015 ATA guideline and radiologist averaged values were calculated (Table 2). The sensitivity, specificity and accuracy of radiologist 1 according to the 2015 ATA guideline 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 difference comparing both radiologists (sensitivity, $p < 0.001$, specificity, $p < 0.001$ and accuracy, $p = 0.003$). 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 : 0.468-0.512) and 0.553 (95% CI : 0.495-0.612), respectively and were significantly different from each other (p-value = 0.038).

The diagnostic performance of both classifier models was derived and compared with the radiologist average values (Table 2). ANN classifier model

showed 74.1% accuracy with sensitivity of 32.3% and specificity of 90.1%. Similarly, SVM classifier model showed 69.0% accuracy, 41.7% sensitivity and 79.4% specificity. Both classifier models showed higher accuracy compared to the radiologist averaged values and with statistical significance in the ANN model (p <.001). The cross validated AUC values of ANN and SVM were 0.646 and 0.599, respectively. The AUC values of ANN and SVM classifier models were 0.612 (95% CI : 0.561-0.662) and 0.605 (95% CI : 0.550-0.661), 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 value of ANN classifier model yielded higher values compared with radiologist 1 and radiologist 2 (p <0.001 and p = 0.085, respectively). The AUC value of SVM classifier model also showed higher values compared with radiologist 1 and radiologist 2 (p < 0.001 and p = 0.146, respectively). Example images of discrepancy cases were shown in Figures 3 and 4.

Table 2. Diagnostic performance of two radiologists, radiologist-averaged value and the classifier models at thyroid nodules on US for differentiating follicular neoplasm

	Rad1	**Rad2**	**Radiologist-averaged value**	**SVM**	**ANN**	***p*** [a]	***p*** [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	<.001
Accuracy	69.5%	60.1%	64.8%	69.0%	74.1%	0.137	<.001

Rad1, radiologist 1; Rad2, radiologist 2; SVM, support vector machine; ANN, artificial neural network

[a] *p*-value obtained by comparing corresponding values with SVM and radiologist-averaged values

[b]*p*-value obtained by comparing corresponding values with ANN and radiologist-averaged values

3. Interobserver variability

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

IV. DISCUSSION

Preoperative US and fine needle aspiration cytology have been used to differentiate benign and malignant thyroid nodules and have given good diagnostic performance in preoperatively distinguishing papillary thyroid cancer [26,27]. However, these methods have limited role in discriminating follicular adenoma and carcinoma of the thyroid gland. Although, certain US features such as solid appearance and hypoechogenicity have been associated with follicular carcinoma compared to follicular adenoma, these features show high sensitivity (ranging 79.8% to 90.1%) and low specificity (ranging 30.8% to 50.0%) in discriminating follicular carcinoma from adenoma [7,28,29]. Similarly, despite the fact that certain subtypes of adenoma such as colloid adenoma may be distinguished from follicular carcinoma by fine needle aspiration cytology, certain subtypes of adenoma such as Hurthle cell adenoma are known to be indistinguishable from follicular carcinoma and thus significant gray zone exists [16,30-32]. Core needle biopsy or intraoperative frozen section discriminate follicular adenoma from carcinoma with slightly high specificity and low sensitivity and frequent indeterminate results hinders their practical use as an independent tool [33-36].

In our study, we developed radiomics-based classifier models based to differentiate follicular adenoma and carcinoma through preoperative US images. The diagnostic performance of our models was evaluated and compared with experienced radiologists. Experienced radiologists categorized each nodule according to the 2015 ATA guideline and 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 compared to experienced radiologists (radiologist-averaged 64.8% *vs.* SVM 69.0% *vs.* ANN 74.0%). Also, 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. Combined with the aforementioned traditional US imaging findings, which showed relatively high sensitivity and low specificity, our radiomics-based classifier models may have additive value in the preoperative discrimination of follicular adenoma and carcinoma. To our knowledge, this is the first study to apply radiomics in preoperative US to predict malignancy in a study population exclusively including follicular neoplasm of the thyroid gland. Several previous studies have applied radiomics to predict malignancy of thyroid nodules on US, but have done so regardless of histologic subtype [37]. Liang *et al*. included 137 thyroid nodules (52 benign and 82 malignant nodules) in their training cohort and developed a formula to calculate radiomics score for each nodule using radiomics features extracted from preoperative US. Similar to our study, 1044 features were initially extracted and then reduced to 19 features using the LASSO regression model. The diagnostic performance of their radiomics score model showed AUC of 0.921 in predicting malignancy and showed better performance compared to experienced and junior radiologists referring to 2017 Thyroid Imaging, Reporting, and Data System scoring criteria [38]. Another study by Yu *et al*.

included 610 thyroid nodules (403 benign and 207 malignant nodules) to predict malignancy. Texture features were extracted from each nodule and were used to train ANN and SVM based classifier models. ANN and SVM models showed 90.0% and 86.0% accuracy in predicting malignancy, respectively [15]. 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 that included all thyroid nodules. However, experienced radiologists in our study also showed lower performance in predicting malignancy compared to these studies. The reason for this may be the difference in conformity of the ultrasound findings in predicting follicular carcinoma [39]. Significant gray zone exists in US findings between follicular adenoma and carcinoma, and therefore, there may be potentially a bigger role for machine learning based classifier model in discriminating follicular adenoma from carcinoma in larger confirmative studies.

Interobserver variability in discriminating malignant to benign thyroid nodules in overall have shown substantial agreement ($\kappa = 0.61 – 0.79$) among experienced radiologists [40-42]. To date, there is no study stating the interobserver variability of US assessment results limited to follicular neoplasm of the thyroid gland on cytology. In our study, the interobserver variability among two radiologists was poor showing only slight agreement ($\kappa = 0.076$), even though both radiologists had more than 10 years of experience on thyroid US. Similarly, all performance variables (sensitivity, specificity and accuracy) in each radiologist showed significantly different values from one another even though the same guideline was used as reference for decision. These findings suggest that US analysis in discriminating follicular adenoma and carcinoma is much more subjective with significant gray zone and thus gives low reproducibility. Therefore, radiomics-based classifier models using quantitative information

from US images have the potential to give more objective results in preoperative discrimination of follicular adenoma and carcinoma on US.

There are several limitations in our study. First, the study population is small, with a total of 348 nodules consisting of 252 follicular adenomas and 96 follicular carcinomas. Due to this small study population, leave-one-out cross validation method was used for model validation rather than subgrouping a separate validation set. 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, gender and tumor size could be predictor factors of malignancy in follicular neoplasm of the thyroid [7,43]. However, in our study, demographic data showed no significant difference between benign and malignant subgroup and thus these variables were not included as input variables to the classifier system. Larger data sets may reveal the potential role of demographic data in diagnosis of follicular neoplasm of the thyroid gland. Third, external validation was not done in our study. Due to the low prevalence of follicular neoplasm and even lower incidence of follicular thyroid carcinoma, it was difficult to prepare a separate group of patient for external validation. Further studies with larger sample size are needed for further validation. Lastly, the diagnostic performance and interobserver variability is questionably low, even though two radiologists have more than 10 years of experience in thyroid imaging. This result may be due to the fact that US differentiation of follicular adenoma and carcinoma is challenging and moreover, majority of the patients in the carcinoma group were minimally invasive type (88/96, 91.6%), which is more indistinguishable from adenoma.

V. CONCLUSION

Our current SVM and ANN classifier models which used texture features as input variables, showed better diagnostic accuracy compared to

experienced radiologists referring to the ATA guideline with high specificity in differentiating follicular carcinoma from adenoma. This study is a preliminary study and with further validation and refinement, classifier models may have the potential to aid attending radiologists in differentiating thyroid follicular neoplasm.

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;235:656-62; discussion 62-4.
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;56:383-9.
4. McHenry CR, Phitayakorn R. Follicular adenoma and carcinoma of the thyroid gland. Oncologist 2011;16:585-93.
5. Yoon JH, Kim E-K, 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. International journal of endocrinology 2014;2014.
6. Kim E-K, Park CS, Chung WY, Oh KK, Kim DI, Lee JT, et al. New sonographic criteria for recommending fine-needle aspiration biopsy of nonpalpable solid nodules of the thyroid. American Journal of Roentgenology 2002;178:687-91.
7. Sillery JC, Reading CC, Charboneau JW, Henrichsen TL, Hay ID, Mandrekar JN. Thyroid follicular carcinoma: sonographic features of 50 cases. American Journal of Roentgenology 2010;194:44-54.
8. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278:563-77.
9. 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. European journal of cancer 2012;48:441-6.
10. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. Radiographics 2017;37:505-15.
11. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine Learning methods for Quantitative Radiomic Biomarkers. Sci Rep 2015;5:13087.
12. 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;15:853-8.
13. 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; doi:10.2214/ajr.15.15813.1-6.
14. 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;35:775-82.
15. 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; doi:10.1007/s00405-017-4562-3.
16. Baloch ZW, Fleisher S, LiVolsi VA, Gupta PK. Diagnosis of "follicular neoplasm": a gray zone in thyroid fine-needle aspiration cytology. Diagn Cytopathol 2002;26:41-4.
17. 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;22:3007-13.
18. 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.
19. Castellano G, Bonilha L, Li LM, Cendes F. Texture analysis of medical images. Clin Radiol 2004;59:1061-9.
20. Liu C, Wechsler H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans Image Process 2002;11:467-76.
21. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 1996:267-88.
22. Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural processing letters 1999;9:293-300.
23. Orr GB, Müller K-R. Neural networks: tricks of the trade: Springer; 2003.
24. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. Biometrics 1997;53:567-78.
25. Kundel HL, Polansky M. Measurement of observer agreement. Radiology 2003;228:303-8.
26. 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;247:762-70.
27. Bongiovanni M, Spitale A, Faquin WC, Mazzucchelli L, Baloch ZW. The Bethesda system for reporting thyroid cytopathology: a meta-analysis. Acta cytologica 2012;56:333-9.
28. Kobayashi K, Fukata S, Miyauchi A. Diagnosis of follicular carcinoma of the thyroid: role of sonography in preoperative diagnosis of follicular nodules. J Med Ultrason (2001) 2005;32:153-8.
29. 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;134:1312-5.
30. Mazzaferri EL. Management of a solitary thyroid nodule. N Engl J Med 1993;328:553-9.
31. Greaves TS, Olvera M, Florentine BD, Raza AS, Cobb CJ, Tsao- Wei DD, et al. Follicular lesions of thyroid. Cancer Cytopathology 2000;90:335-41.
32. Caraway NP, Sneige N, Samaan NA. Diagnostic pitfalls in thyroid fine-needle aspiration: a review of 394 cases. Diagn Cytopathol 1993;9:345-50.
33. Yoon RG, Baek JH, Lee JH, Choi YJ, Hong MJ, Song DE, et al. Diagnosis of

thyroid follicular neoplasm: fine-needle aspiration versus core-needle biopsy. Thyroid 2014;24:1612-7.

34. 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.
35. 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;11:94-8.
36. Leteurtre E, Leroy X, Pattou F, Wacrenier A, Carnaille B, Proye C, et al. Why do frozen sections have limited value in encapsulated or minimally invasive follicular carcinoma of the thyroid? Am J Clin Pathol 2001;115:370-4.
37. Sollini M, Cozzi L, Chiti A, Kirienko M, ejr. 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.
38. 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;28:1024-33.
39. 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;8:192-7.
40. Choi SH, Kim EK, Kwak JY, Kim MJ, Son EJ. Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. Thyroid 2010;20:167-72.
41. 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;42:51-9.
42. 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.
43. Paramo JC, Mesko T. Age, tumor size, and in-office ultrasonography are predictive parameters of malignancy in follicular neoplasms of the thyroid. Endocr Pract 2008;14:447-51.

APPENDICES

Appendix 1. 96 extracted features from US images of the thyroid nodules.

Group		Num.	Features
Densitometric(Histogram)		9	Min, Max, Mean, Stdev, Variance, Skewness, Kurtosis, Energy, Entropy
Morphormetric		13	Area, Roundness, Extent, Compactness, Convex area, Convex perimeter, Roughness, Concavity, Solidity, Feret diameter, Elongation, Wadells circularity, Heywood circularity factor
Texture	GLCM	32	Contrast, Dissimilarity, Homogeneity, ASM, Energy, Probability max, Entropy, Correlation
	GLRLM	16	GLN, RLN, LGRE, HGRE
	Gabor	12	O1S1, O2S1, O3S1, O4S1, O1S2, O2S2, O3S2, O4S2, O1S3, O2S3, O3S3, O4S3
	Haar wavelet, Lv.2	14	Energy LL2, Energy LH2, Energy HL2, Energy HH2, Energy LH1, Energy HL1, Energy HH1, Entropy LL2, Entropy LH2, Entropy HL2, Entropy HH2, Entropy LH1, Entropy HL1, Entropy HH1

ABSTRACT(IN KOREAN)

초음파 영상의 기계 학습을 이용한 갑상선 여포성 종양의 감별진단

<지도교수 곽진영>

연세대학교 대학원 의학과

신일아

목적 : 현재 영상의학 분야에서 많이 이용되고 있는 기계 학습을 이용하여 여포성 샘종과 여포암을 수술 전 초음파를 이용하여 감별해 보고자 한다.

방법 : 본 후향적 연구는 2개의 3차 병원에서 모은 340명의 환자의 348개의 갑상선 결절의 수술 전 초음파 영상을 대상으로 시행되었다. 각 초음파 영상의 그려진 관심영역에서 96개의 라디오믹스 소견을 추출하였으며 그 중 최종적으로 10개의 소견만 라디오믹스 분류 모델의 학습에 이용되었다. 또한, 두 명의 영상의학과 의사도 같은 초음파 영상을 독립적으로 평가하여 2015년도 American Thyroid Association 지침에 따라 각 결절을 양성인지 악성인지 분류하였다. 두 영상의학과 의사와 라디오믹스 기반의 분류 모델들의 진단능을 비교 평가하였

다.

결과 : 245명의 환자로부터 진단된 252개의 여포성 샘종과 95명의 환자에서 진단된 96개의 여포암이 본 연구에 포함이 되었다. 두 영상의학과의 평균 민감도, 특이도 그리고 정확도는 24.0%, 84.0% 그리고 64.8% 였다. 인공신경망에 기반한 분류 모델과 Support vector machine에 기반한 분류 모델의 민감도, 특이도 정확도는 각각 32.3%, 90.1%, 74.1% 그리고 41.7%, 79.4%, 69.0% 였다. 두 영상의학과 의사간의 카파 (κ) 값은 0.076 으로 경미한 상관 관계만을 보였다.

결론 : 기계 학습에 기반한 분류 모델은 영상의학과 의사들보다 수술 전 초음파 영상을 이용한 여포성 샘종과 여포함의 감별에 더 나은 진단능을 보였으며, 이러한 분류 모델들은 상관 관계가 높지 못한 사람과 다르게 비교적 일정한 결과를 보여줄 수 있는 데에 의의가 있겠다.

핵심되는 말 : 라디오믹스, 인공신경망, 서포트 벡터 머신, 여포성 샘종, 갑상선 여포암