Research Article

Diagnostic Value of Machine Learning-Based Quantitative Texture Analysis in Differentiating Benign and Malignant Thyroid Nodules

Bulent Colakoglu,[1] Deniz Alis ,[2] and Mert Yergin[3]

[1]*Vehbi Koç Foundation American Hospital, Department of Radiology, Istanbul, Turkey*
[2]*Istanbul Mehmet Akif Ersoy Thoracic and Cardiovascular Surgery Training and Research Hospital, Department of Radiology, Halkali, Istanbul, Turkey*
[3]*Bahcesehir University, Department of Software Engineering and Applied Sciences, Istanbul, Turkey*

Correspondence should be addressed to Deniz Alis; drdenizalis@gmail.com

Received 11 August 2019; Revised 3 October 2019; Accepted 4 October 2019; Published 31 October 2019

Academic Editor: Francesca De Felice

Aim. The aim of this study is to evaluate the diagnostic value of machine learning- (ML-) based quantitative texture analysis in the differentiation of benign and malignant thyroid nodules. *Materials and methods*. A sum of 306 quantitative textural features of 235 thyroid nodules (102 malignant, 43.4%; 133 benign, 56.4%) of a total of 198 patients were investigated using the random forest ML classifier. Feature selection and dimension reduction were conducted using reproducibility testing and a wrapper method. The diagnostic accuracy, sensitivity, specificity, and area under curve (AUC) of the proposed method were compared with the histopathological or cytopathological findings as reference methods. *Results*. Of the 306 initial texture features, 284 (92.2%) showed good reproducibility (intraclass correlation ≥0.80). The random forest classifier accurately identified 87 out of 102 malignant thyroid nodules and 117 out of 133 benign thyroid nodules, which is a diagnostic sensitivity of 85.2%, specificity of 87.9%, and accuracy of 86.8%. The AUC of the model was 0.92. *Conclusions*. Quantitative textural analysis of thyroid nodules using ML classification can accurately discriminate benign and malignant thyroid nodules. Our findings should be validated by multicenter prospective studies using completely independent external data.

1. Introduction

Thyroid nodules are common, with a prevalence of up to 67% in the adult population [1, 2]. Approximately 5%–15% of these nodules are malignant, and the differentiation of malignant and benign nodules is mandatory for forming individual management strategies [3–7]. Ultrasound (US) is the first and most commonly used imaging modality for the evaluation of thyroid nodules [2, 3]. The nodules that are strongly suspected to be malignant as appraised by US are further evaluated by fine-needle aspiration biopsy (FNAB) or tissue biopsies; hence, a noninvasive method with an ability to differentiate malignant and benign nodules is desirable [8, 9]. Sonographic features such as irregular margin, solid composition, hypoechogenicity, elongated shape, and microcalcifications indicate malignancy [8, 9]. However, these features are somewhat qualitative, and the experience of a radiologist has a substantial influence on diagnostic accuracy [10–15]. Over the years, various reporting systems have been introduced to reduce inconsistencies among radiologists and promote communication between clinicians and radiologists. However, these reporting systems still rely on a radiologist's subjective interpretations. Moreover, some radiologists are reluctant to use these reporting systems owing to their complexity [8–10].

Radiomics is defined as the machine learning- (ML-) or deep learning-based mining of quantitative texture features extracted from conventional imaging modalities. The aim is to improve the precision and diagnostic accuracy of imaging

methods, mostly in the field of cancer research [16, 17]. Several studies have demonstrated the feasibility of ML-based texture features in differentiating between benign and malignant thyroid nodules [18–28]. Nevertheless, ML-based texture analysis for the evaluation of thyroid nodules is still in its infancy, and further research is necessitated.

Herein, we evaluate the diagnostic performance of ML-based quantitative texture analysis in differentiating malignant and benign thyroid nodules.

2. Materials and Methods

The local ethics committee approved this retrospective study, which was conducted between January 2015 and January 2019. We reviewed our picture archive and communicating system to identify patients who underwent thyroid US examination for thyroid nodules. The patients in whom the thyroid nodules were determined as benign or malignant according to FNAB or surgical pathology were included in the study. Nodules that fell into the non-diagnostic or indeterminate categories according to the Bethesda Classification System and thyroid nodules <1 cm in diameter were excluded from the study [4]. All of the nodules were evaluated by the same radiologist (B.C.), who had more than 20 years of experience in thyroid US, with the same device (LOGIQ E9 with XDclear, General Electric (GE) Healthcare, Wauwatosa, WI, USA) using a linear array transducer (ML6-15) with a frequency range of 12 to 15 MHz. Grayscale images of the thyroid nodules in the axial plane were selected for further analysis.

2.1. Texture Feature Extraction. US images contain inherent impulse and salt-and-pepper type noise; hence, an anisotropic median filter was applied to all US images before the texture extraction step to remove noise while preventing the edges and boundaries of images being blurred. QMaZda texture analysis software was used for quantitative texture feature extraction [29]. The radiologists manually delineated the borders of the thyroid nodules for texture extraction. The image histogram was remapped within $\pm 3\sigma$ of the grayscale levels to prevent texture features from being affected by image characteristics such as contrast or brightness [29].

A total of 306 texture features were extracted for further analysis [28]: first-order histograms (13), gradient-map-based features (5), gray-level co-occurrence matrix (GLCM) features (176), gray-level run-length matrix (GRLM) features (28), autoregressive model features (5), Haar wavelets (12), Gabor transform features (24), histogram of oriented gradients (HOG; 8), and local binary patterns (LBP; 35) [30]. GLCM and GRLM were calculated at 5 bits per pixel, and gradient-map-based features were calculated at 4 bits per pixel. First-order histograms, autoregressive model features, and Haar wavelet features were calculated at 8 bits per pixel. LBPs were calculated by one of the three algorithms:over-complete (Oc), transition (Tr), and center-symmetric (Cs),with respect to the number of $4n$ neighbors. The extracted texture features were further entered into ML-based analyses.

2.2. Feature Selection and Dimension Reduction. Waikato Environment for Knowledge Analysis (WEKA) toolkit version 3.8.2 (University of Waikato, Hamilton, New Zealand) was used for feature selection [31]. The current study, similar with most other studies involving quantitative texture analysis, had a substantially higher number of texture features ($n = 306$) than thyroid nodules in the study cohort ($n = 235$). This is a recognized problem in ML analysis and is also known as the curse of dimensionality. As a consequence, there is a significant risk of overfitting the model. Over-fitted models can be briefly described as models that strictly fit the training data but show poor performance on new cases, namely, the test set. Hence, there is a wide agreement as to the importance of feature selection before creating the ML model for classification or prediction tasks. To detect the most appropriate features while discarding irrelevant ones for the model, we employed three consecutive steps. First, two radiologists (B.C. and D.A.) drew regions of interest onto randomly selected images of benign and malignant thyroid nodules (10 images each) to assess the reproducibility of the extracted texture features. Features with a good intraclass correlation (ICC) value (≥0.80) were further considered in the following steps. Next, a scheme-dependent feature selection method, a wrapper subset evaluation using 10-fold cross-validation, was applied [31, 32].

A wrapper method is a supervised scheme-dependent feature selection technique that evaluates the features according to their importance to the model [32]. The wrapper method evaluates the relevance of the attributes based on a classifier, which was the random forest classifier in the present work. The wrapper method first creates multiple subsets of the features and then tests the performance of these subsets to find the best combination of features. There are several types of wrapper methods depending on the search method. The current study uses the wrapper method with linear forward stepwise selection, in which the search for the most relevant features for the model begins with a null model and continues until the best subset of the attributes are determined [32]. We applied feature selection after cross validation; hence, relevant features were selected using only the training partitions of the dataset to avoid the "double-dipping" phenomenon, which occurs when the whole dataset is used for the selection and might lead to biased or over-optimistic results [32, 33]. The selected features were further processed using the ML classifier (a random forest) to assess its diagnostic performance in discriminating benign and malignant thyroid nodules. Figure 1 summarizes the pipeline of the present work.

2.3. ML Models and Statistical Analyses. WEKA toolkit version 3.8.2 was used to develop the ML model and to test its performance. Only selected quantitative texture features were used in the ML analysis. The present study used the random forest classifier to build the ML model. The random forest is a well-known ML algorithm for classification tasks and has an inherent resistance to overfitting [34]. The random forest is an ensemble learning method. It chooses

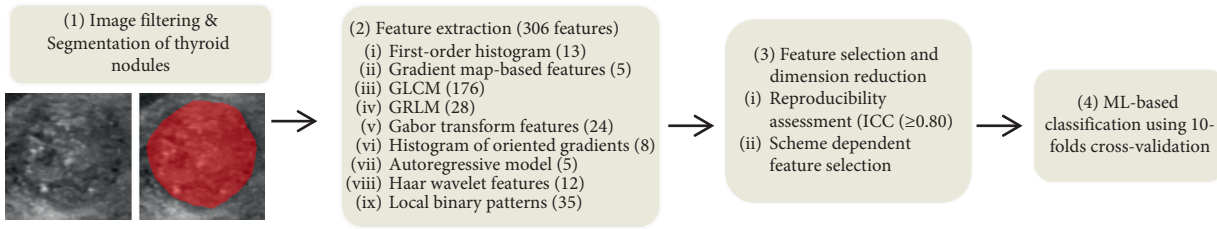

Figure 1: The scheme summarized the main workflow of the current study.

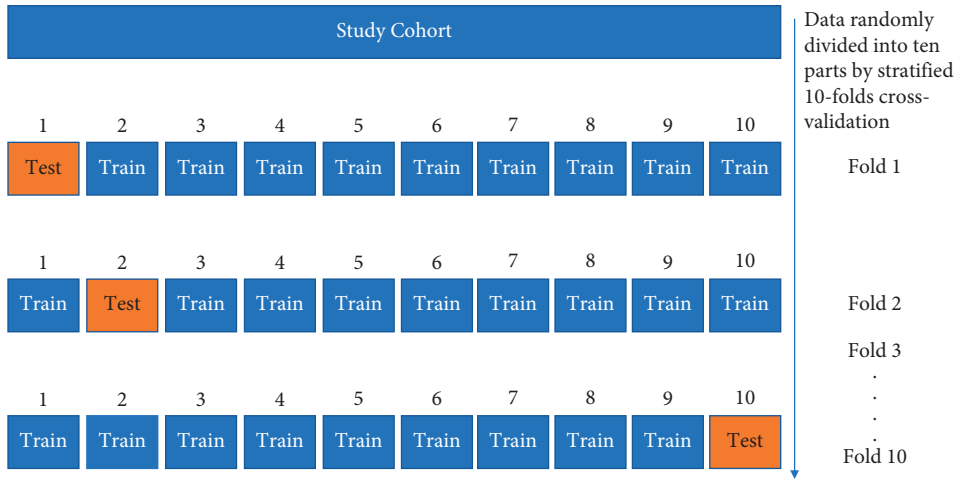

Figure 2: Evaluation of the model's performance by 10-fold cross-validation. 10-fold cross-validation first randomly divides all the data into ten parts then holds out 10% of the data for testing. This process is repeated ten times, and then the mean accuracy for the algorithm is calculated.

random data points from the dataset to build multiple decision trees and then uses all of these decision trees to improve the performance of the final prediction [31]. We did not split the cohort into training and testing groups; instead, we applied stratified 10-fold cross-validation, which randomly divides all the data into ten parts and then holds out 10% of the data for testing. This process is repeated ten times [35]. A detailed illustration of the 10-fold cross-validation method is given in Figure 2.

The diagnostic performance of the ML model was assessed using correlation matrices, which shows the results as the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) according to the histopathological sampling results. The following formulas were then used to determine performance: sensitivity = TP/(TP + FN), specificity = TN/(TN + FP), and diagnostic accuracy = (TP + TN)/(TP + TN + FP + FN). The receiver operating curve was drawn for the ML model and the area under the curve (AUC) was calculated. The wrapper method is able to identify the best subset for the ML model, but it does not provide further information regarding the relative importance of the selected features. Hence, we employed an information gain attribute evaluator, which evaluates the worth of an attribute (in this case, the discriminative power of the attributes for benign and malignant thyroid nodules) by measuring the information gain with respect to the class [31]. The following formula was used for the information gain attribute evaluator: InfoGain(Class, Attribute) = $H(\text{Class}) - H(\text{Class} \mid \text{Attribute})$, where H represents the amount of information in a unit called bits and ranges in value between 0 and 1 [31]. The information value increases as the value approaches 1.

3. Results

The final cohort study comprised a total of 235 thyroid nodules of 198 patients (150 females, 48 males; age range 18–81 years; and mean age 44.55 years). There were 98 patients with 102 malignant thyroid nodules and 100 patients with 133 benign thyroid nodules. Of the 98 patients with malignant thyroid nodules, 33 were male and 65 were female with a mean age of 42.12 ± 14.55 years. Of the 100 patients with benign thyroid nodules, 22 were male and 78 were female with a mean age of 46.35 ± 17.12 years. Among the 102 malignant nodules, the FNAB results of 82 nodules (80.3%) were also confirmed by surgical pathology as 73 papillary thyroid carcinomas (89%) and nine follicular variants of papillary cancer (11%). The other malignant thyroid nodules ($n = 20$) had only a cytopathological diagnosis because the patients did not undergo an operation at our institution. Of the 306 initial texture features, 284 (92.2%) showed good reproducibility (ICC ≥ 0.80) and were further used for the ML-based evaluation. A total of seven texture features were selected for the final model: one histogram (HistPerc 99), one HOG (HogO8b2), four GRLM (GrlmHRLNonUni, GrlmHMGLevNonUni, GrlmNRLNonUni, and GrlmZRLNonUni), and one GLCM (GlcmZ3AngScMom).

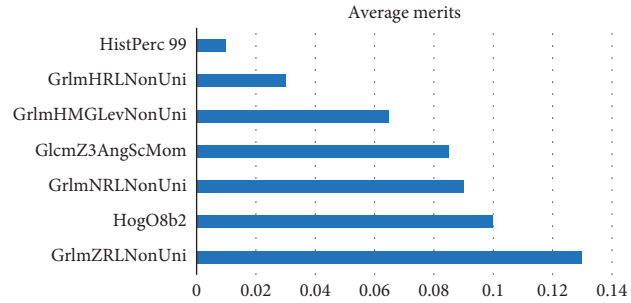

Figure 3: A total of seven texture features were selected for the final model: one histogram (HistPerc 99), one HOG (HogO8b2), four GRLM (GrlmHRLNonUni, GrlmHMGLevNonUni, GrlmNRLNonUni, and GrlmZRLNonUni), and one GLCM (GlcmZ3AngScMom). The information gain attribute evaluator identified that GrlmZRLNonuni was the most important feature in the final model followed by HogO8b2 and GrlmNRLNonUni. The formula of the information gain attribute evaluator was InfoGain(Class, Attribute) = H(Class) − H(Class | Attribute), where H represents the amount of information in a unit called bits and ranges in value between 0 and 1.

HistPerc 99 is an alternative to maximum intensity that is equal to the highest intensity in the defined region of interest [36, 37]. The mean HistPerc 99 values of the benign nodules were higher than those of the malignant ones in the present study, indicating the higher intensity values in benign tumors. The GRLM counts the runs of pixels with the same gray level in different directions [36, 37]. The selected GRLM features were the following in the present work: GrlmHRLNonUni, GrlmHMGLevNonUni, GrlmNRLNonUni, and GrlmZRLNonUni. In the feature names, "Grlm" represents GRLM, the letters "H," "N," and "Z" represent the direction of that feature: $H = 0^{\circ}$ (horizontal), $Z = 45^{\circ}$, and $N = 135^{\circ}$ [29]. Furthermore, MGLevNonUni and RLNonUni represent nonuniformity, in which a higher value indicates heterogeneity [29, 36, 37]. In the present work, malignant thyroid nodules had more runs with nonuniform values, indicating the heterogeneity of the nodules. GLCM features count the number of co-occurrences of pixels with specified gray-levels. Pairs of pixels are considered, such that one of the pixels is situated at an offset (Δx, Δy) with respect to the other [36, 37]. The random forest model used GlcmZ3AngScMom, where "Glcm" represents the GLCM, "Z3" represents the direction (45°) and the offset of the pixels, and AngScMom represents angular second movement and also energy, which is a measure of homogeneous patterns in the image [29, 36, 37]. Benign thyroid nodules had higher mean GlcmZ3AngScMom values in the present work. The HOG counts the number of occurrences of gradient. It identifies a sudden change in the pixel values, which is called the gradient [29, 36, 37]. A positive gradient refers to a change from a lower to higher pixel value while a negative gradient refers to a higher-to-lower change in value. HOGs are also a known marker of heterogeneity. In the present work, malignant nodules had higher gradient reflecting the heterogeneity of the tumor.

The random forest classifier accurately identified 87 of the 102 malignant thyroid nodules and 117 of the 133 benign thyroid nodules. These values equate to a diagnostic sensitivity of 85.2%, a specificity of 87.9%, and an accuracy of 86.8%. The AUC was calculated as 0.92 for the model. The average values of the quantitative texture features for the differentiation of malignant and benign thyroid nodules are shown in Figure 3.

4. Discussion

The present study demonstrated that the ML model using the random forest classifiers with selected texture features can successfully discriminate malignant and benign thyroid nodules. The selected texture features of the random forest model consisted of histogram, HOG, GRLM, and GLCM features. The selected second-order features mainly reflect the increased heterogeneity of the malignant thyroid nodules, whereas the histogram feature represents the hypoechogenic characteristic of the malignant nodules.

4.1. Related Work. Several authors have evaluated the diagnostic value of ML-based texture analysis for the differentiation of benign and malignant thyroid nodules. The diagnostic accuracy has even reached 100% in some of these works. For instance, Acharya et al. [19–21] demonstrated that ML-based texture analysis had a diagnostic accuracy ranging from 98.3% to 100%, but the study cohorts consisted of only 20 nodules in three of their reports. In the work by Chang et al. [22], the support vector machines classifier showed a diagnostic accuracy reaching up to 98.3% for 59 nodules. A recent work by Prochazka et al. [38] employed a random forest and support vector machine classifier for evaluating segmentation-based fractal texture analysis, and the authors achieved a diagnostic accuracy of 94.3% with their model. Although all studies mentioned above have yielded promising results, the use of a small sample size in ML-based diagnostic models will undoubtedly introduce bias and variance [39]. Furthermore, using such small cohorts increases the risk of overfitting and limits the generalizability of the results [39].

To our knowledge, there are few other studies on the ML-based quantitative texture analysis of thyroid nodules that have a cohort size that is comparable to the size of the one in our work. These works reported diagnostic accuracies ranging from 78.5% to 94.3%, which is comparable with the diagnostic accuracy of 86.8% obtained by the random forest in the present work [23, 26, 28]. Song et al. [26] evaluated 16 GLCM features using logistic regression, artificial neural network, random forest, boost, SVM, and random tree models. They found that the logistic regression model achieved the highest diagnostic accuracy. Similar to our work, they did not use different data for training and testing; instead, they implemented 10-fold cross-validation [26]. Acharya et al. [23] evaluated the Gabor transform features of 242 benign and malignant thyroid nodules using support vector machines, k-nearest neighbors, multilayer perceptron, and C4.5 decision tree classifiers. In their work, the C4.5 decision tree classifier achieved the best diagnostic performance with a diagnostic accuracy of 94.3%. Yu et al.

[28] implemented artificial neural network classifiers for the evaluation of two morphological and 65 different texture features. They trained the initial models using images of 610 thyroid nodules with 10-fold cross-validation and achieved 99% diagnostic accuracy. They externally validated their model using images of 50 thyroid nodules, which resulted in a diagnostic accuracy of 90% [28]. We suggest that the better accuracy obtained by Yu et al. [28] might be a consequence of the inclusion of semantic parameters such as the orientation and boundaries of the nodules. It is well known that semantic features such as vertical shape and irregular borders are closely associated with malignancy; hence, they might improve the ML-based models.

4.2. Limitations. First and foremost, there is an inevitable selection bias in the present work because we only included thyroid nodules categorized as benign or malignant according to FNAB or surgical pathology; hence, we excluded most of the benign nodules that did not have biopsy results. In daily practice, a relatively small number of thyroid nodules are scheduled for histopathological examination, and most of the thyroid nodule data with benign features were follow-ups using US [3, 4]. Therefore, benign thyroid nodules in the present work might not cover all types of benign nodules. Second, we did not evaluate the semantic features of thyroid nodules nor integrate qualitative US features such as echogenicity or nodule composition because our aim was to assess the diagnostic value of textural analysis alone. Third, we neither compared the diagnostic accuracy of our ML model with human evaluators nor evaluated the diagnostic accuracy of a human evaluator supplemented by the ML model. Hence, we suggest that further studies investigating the diagnostic accuracies of human evaluators and ML-based classifications as well as the assistive value of ML-based models for human evaluators might be worthwhile. Fourth, we did not use separate test and training groups; instead, we implemented the 10-fold cross-validation algorithm, which allows us to use the same cohort as test and training subjects [35]. Finally, although it is not peculiar to the present work, to date, many ML-based classification systems and an abundant number of different texture features have been evaluated for differentiating thyroid nodules as benign or malignant. Hence, the standardization of the models and evaluated features is a concerning issue that prevents the use of ML models for the characterization of thyroid nodules in practice [40].

5. Conclusion

We demonstrated that an ML classifier, the random forest, with selected textural features can achieve 85.2% sensitivity, 87.9% specificity, and 86.8% diagnostic accuracy with an AUC of 0.92 in the task of differentiating malignant thyroid nodules from benign ones. The texture features selected in this study indicate that malignant thyroid nodules have increased heterogenicity and lower echogenicity than benign thyroid nodules. We acknowledge that our findings should be validated by prospective multicenter studies using a completely independent external dataset.

Abbreviations

AUC: Area under the curve
GLCM: Gray-level co-occurrence matrix
GRLM: Gray-level run-length matrix
HOG: Histogram of oriented gradients
ML: Machine learning
US: Ultrasound
WEKA: Waikato Environment for Knowledge Analysis.

Data Availability

The data supporting our findings can be found in the article. Given the strict regulations of our institutional review board, further data of the patients are only available from the corresponding author on reasonable request.

Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Institutional Review Board approval was obtained.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] P. A. Singer, "Evaluation and management of the solitary thyroid nodule," *Otolaryngologic Clinics of North America*, vol. 29, no. 4, pp. 577–591, 1996.

[2] S. Guth, U. Theune, J. Aberle, A. Galach, and C. M. Bamberger, "Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination," *European Journal of Clinical Investigation*, vol. 39, no. 8, pp. 699–706, 2009.

[3] B. R. Haugen, E. K. Alexander, K. C. Bible 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*, vol. 26, no. 1, pp. 1–133, 2016.

[4] D. S. Cooper, G. M. Doherty, B. R. Haugen et al., "Revised American thyroid association management guidelines for patients with thyroid nodules and differentiated thyroid cancer," *Thyroid*, vol. 19, no. 11, pp. 1167–1214, 2009.

[5] E. E. Werk Jr., B. M. Vernon, J. J. Gonzalez, P. C. Ungaro, and R. C. McCoy, "Cancer in thyroid nodules. A community hospital survey," *Archives of Internal Medicine*, vol. 144, no. 3, pp. 474–476, 1984.

[6] M. C. Frates, C. B. Benson, P. M. Doubilet et al., "Prevalence and distribution of carcinoma in patients with solitary and multiple thyroid nodules on sonography," *The Journal of Clinical Endocrinology & Metabolism*, vol. 91, no. 9, pp. 3411–3417, 2006.

[7] J. Jin and C. R. McHenry, "Thyroid incidentaloma," *Best Practice & Research Clinical Endocrinology & Metabolism*, vol. 26, no. 1, pp. 83–96, 2012.

[8] C. C. Reading, J. W. Charboneau, I. D. Hay, and T. J. Sebo, "Sonography of thyroid nodules: a "classic pattern" diagnostic approach," *Ultrasound Quarterly*, vol. 21, no. 3, pp. 157–165, 2005.

[9] F. N. Tessler, W. D. Middleton, E. G. Grant et al., "ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee," *Journal of the American College of Radiology*, vol. 14, no. 5, pp. 587–595, 2017.

[10] G. Grani, M. D'Alessandri, G. Carbotta et al., "Grey-scale analysis improves the ultrasonographic evaluation of thyroid nodules," *Medicine*, vol. 94, no. 27, Article ID e1129, 2015.

[11] S. H. Choi, E.-K. Kim, J. Y. Kwak, M. J. Kim, and E. J. Son, "Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules," *Thyroid*, vol. 20, no. 2, pp. 167–172, 2010.

[12] C. S. Park, S. H. Kim, S. L. Jung et al., "Observer variability in the sonographic evaluation of thyroid nodules," *Journal of Clinical Ultrasoundd*, vol. 38, pp. 287–293, 2010.

[13] S. H. Kim, C. S. Park, S. L. Jung et al., "Observer variability and the performance between faculties and residents: US criteria for benign and malignant thyroid nodules," *Korean Journal of Radiology*, vol. 11, no. 2, pp. 149–155, 2010.

[14] A. Scorza, G. Lupi, S. A. Sciuto, F. Bini, and F. Marinozzi, "A novel approach to a phantom based method for maximum depth of penetration measurement in diagnostic ultrasound: a preliminary study," in *Proceedings of the IEEE International Symposium on MeMea*, Torino, Italy, May 2015.

[15] P. Trimboli, F. Bini, M. Andrioli et al., "Analysis of tissue surrounding thyroid nodules by ultrasound digital images," *Endocrine*, vol. 48, no. 2, pp. 434–438, 2015.

[16] G. D. Tourassi, "Journey toward computer-aided diagnosis: role of image texture analysis," *Radiology*, vol. 213, no. 2, pp. 317–320, 1999.

[17] J. J. M. van Griethuysen, A. Fedorov, C. Parmar et al., "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017.

[18] J. Ding, H. Cheng, C. Ning, J. Huang, and Y. Zhang, "Quantitative measurement for thyroid cancer characterization based on elastography," *Journal of Ultrasound in Medicine*, vol. 30, no. 9, pp. 1259–1266, 2011.

[19] U. R. Acharya, O. Faust, S. V. Sree, F. Molinari, R. Garberoglio, and J. S. Suri, "Cost-effective and non-invasive automated benign & malignant thyroid lesion classification in 3D contrast-enhanced ultrasound using combination of wavelets and textures: a class of ThyroScan™ algorithms," *Technology in Cancer Research & Treatment*, vol. 10, no. 4, pp. 371–380, 2011.

[20] U. R. Acharya, O. Faust, S. V. Sree, F. Molinari, and J. S. Suri, "ThyroScreen system: high resolution ultrasound thyroid image characterization into benign and malignant classes using novel combination of texture and discrete wavelet transform," *Computer Methods and Programs in Biomedicine*, vol. 107, no. 2, pp. 233–241, 2012.

[21] U. R. Acharya, S. Vinitha Sree, M. M. Krishnan, F. Molinari, R. Garberoglio, and J. S. Suri, "Non-invasive automated 3D thyroid lesion classification in ultrasound: a class of ThyroScan™ systems," *Ultrasonics*, vol. 52, no. 4, pp. 508–520, 2012.

[22] Y. Chang, A. K. Paul, N. Kim et al., "Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: a comparison with radiologist-based assessments," *Medical Physics*, vol. 43, no. 1, pp. 554–567, 2016.

[23] U. R. Acharya, P. Chowriappa, H. Fujita et al., "Thyroid lesion classification in 242 patient population using Gabor transform features from high resolution ultrasound images," *Knowledge-Based Systems*, vol. 107, pp. 235–245, 2016.

[24] S.-J. Chen, S.-N. Yu, J.-E. Tzeng et al., "Characterization of the major histopathological components of thyroid nodules using sonographic textural features for clinical diagnosis and management," *Ultrasound in Medicine & Biology*, vol. 35, no. 2, pp. 201–208, 2009.

[25] S. D. Kale and K. M. Punwatkar, "Texture analysis of thyroid ultrasound images for diagnosis of benign and malignant nodule using scaled conjugate gradient back- propagation training neural network," *International Journal of Computer and Engineering Management (IJCEM)*, vol. 16, pp. 33–38, 2013.

[26] G. Song, F. Xue, and C. Zhang, "A model using texture features to differentiate the nature of thyroid nodules on sonography," *Journal of Ultrasound in Medicine*, vol. 34, no. 10, pp. 1753–1760, 2015.

[27] A. A. Ardakani, A. Gharbali, and A. Mohammadi, "Classification of benign and malignant thyroid nodules using wavelet texture analysis of sonograms," *Journal of Ultrasound in Medicine*, vol. 34, no. 11, pp. 1983–1989, 2015.

[28] Q. Yu, T. Jiang, A. Zhou, L. Zhang, C. Zhang, and P. Xu, "Computer-aided diagnosis of malignant or benign thyroid nodes based on ultrasound images," *European Archives of Oto-Rhino-Laryngology*, vol. 274, no. 7, pp. 2891–2897, 2017.

[29] P. M. Szczypiński, M. Strzelecki, A. Materka, and A. Klepaczko, "MaZda—a software package for image texture analysis," *Computer Methods and Programs in Biomedicine*, vol. 94, pp. 66–76, 2009.

[30] G. Thibault, B. Fertil, C. Navarro et al., "Texture indexes and gray level size zone matrix application to cell nuclei classification," in *Proceedings of the 10th International Conference on Pattern Recognition and Information Processing*, pp. 140–145, Barcelona, Spain, July 2009.

[31] E. Frank, A. M. Hall, and I. H. Witten, *Data Mining: Practical Machine Learning Tools and Technique*, Morgan Kaufmann, Burlington, MA, USA, Fourth edition, 2016.

[32] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[33] B. Mwangi, T. S. Tian, and J. C. Soares, "A review of feature reduction techniques in neuroimaging," *Neuroinformatics*, vol. 12, no. 2, pp. 229–244, 2014.

[34] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[35] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[36] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, "Image biomarker standardisation initiative," 2019, https://arxiv.org/abs/1612.07003.

[37] P. Lambin, R. T. H. Leijenaar, T. M. Deist et al., "Radiomics: the bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, vol. 14, no. 12, pp. 749–762, 2017.

[38] A. Prochazka, S. Gulati, S. Holinka, and D. Smutek, "Classification of thyroid nodules in ultrasound images using direction-independent features extracted by two-threshold binary decomposition," *Technology in Cancer Research & Treatment*, vol. 18, Article ID 1533033819830748, 2019.

[39] T. W. Way, B. Sahiner, L. M. Hadjiiski, and H.-P. Chan, "Effect of finite sample size on feature selection and classification: a simulation study," *Medical Physics*, vol. 37, no. 2, pp. 907–920, 2010.

[40] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.