Toward Automated Pre-Biopsy Thyroid Cancer Risk Estimation in Ultrasound

Alfiia Galimzianova, PhD, Sean M. Siebert, BS, Aya Kamaya, MD, Terry S. Desser, MD*, Daniel L. Rubin, MD, MS*
Stanford University School of Medicine, Stanford, CA, USA

Abstract

We propose a computational framework for automated cancer risk estimation of thyroid nodules visualized in ultrasound (US) images. Our framework estimates the probability of nodule malignancy using random forests on a rich set of computational features. An expert radiologist annotated thyroid nodules in 93 biopsy-confirmed patients using semantic image descriptors derived from standardized lexicon. On our dataset, the AUC of the proposed method was 0.70, which was comparable to five baseline expert annotation-based classifiers with AUC values from 0.72 to 0.81. Moreover, the use of the framework for decision making on nodule biopsy could have spared five out of 46 benign nodule biopsies at no cost to the health of patients with malignancies. Our results confirm the feasibility of computer-aided tools for noninvasive malignancy risk estimation in patients with thyroid nodules that could help to decrease the number of unnecessary biopsies and surgeries.

Introduction

The incidence of thyroid nodules, both benign and malignant, has been consistently increasing in the United States in the recent decades. Much of this increase, it is believed, is due to increased utilization of imaging[1], with concomitant increased detection of asymptomatic thyroid nodules. In fact, autopsy studies report that up to 50-67% of adults nationwide are expected to have thyroid nodules[2,3], while only 0.2% of the population is reported to have thyroid cancer[4]. Current definitive diagnosis of thyroid nodules requires tissue biopsy or even surgery, while only 5-7% of these nodules are found to be malignant[5]. This inevitably exposes the majority of the patients to unnecessary health risks associated with these invasive tests and increases societal healthcare costs substantially. Therefore, there is a critical need for methods to reliably estimate the malignancy risks of thyroid nodules to decrease the number of invasive interventions being performed in low-risk benign nodules. Intensive research has been underway in the radiology, endocrinology, and surgery communities to attempt to identify those patients at high risk and who merit invasive diagnostic intervention, yet despite this, accurate diagnosis based on imaging findings remains very challenging.

Ultrasound imaging (US) is the standard-of-care imaging modality used to visually assess the risks of malignancy, define the necessity of biopsy-based definitive diagnosis and guide the fine-needle aspiration biopsy. In its recent guidelines for 2015[6], the American Thyroid Association recommends biopsy for thyroid nodules at size thresholds specified for five nodule appearance patterns associated with different risks of malignancy. However, such pattern-oriented approaches are not collectively exhaustive descriptors of nodule appearance in US and thus are not able to provide any scores to some nodules[7]. Therefore, there is a need for methods that provide comprehensive evaluation of thyroid nodules based on collectively exhaustive sets of US features. Several Thyroid Imaging Reporting and Data System (TIRADS) classification systems have been proposed over the last years with the aim of providing a systematic approach evaluating the cancer risks of thyroid nodules based on multiple US features. The newly developed TIRADS lexicon from the American College of Radiology (ACR)[8] provides a set of standardized US imaging descriptors (visual semantic features) used in expert annotation-based scoring systems[9–13] (TIRADS classifiers) to estimate the risk that a nodule is a cancer before biopsy. However, extraction of these qualitative features imposes additional burden on radiologists and is subject to intra- and interrater variability. Moreover, the scoring systems for nodule malignancy are often oversimplified due to their design for human use. Automated analysis of nodule properties by extraction of computational features from US images and estimation of the risks using machine learning based classifiers could help reduce the expert labor and eliminate the associated variability.

*These authors contributed equally to this work

In this work, we present a computational framework for automated cancer risk estimation from US images, which is based on analysis of a rich set of computational features using a random forest classifier (RF). Unlike existing TIRADS classifiers, the proposed framework does not require qualitative assessments from the expert radiologist. We validated the framework on a dataset of US images of cases with biopsy-proven diagnosis. Our proposed framework compares favorably to other methods and has potential value for computer-aided diagnosis to help decrease the high number of biopsies in patients ultimately found to have benign disease, while identifying those patients at risk of thyroid cancer.

Materials and Methods

Datasets

Ultrasound imaging data of patients that underwent thyroid nodule examination at Stanford Hospital from year 2010 to 2015 were collected retrospectively with the approval by the institutional review board (IRB). Our final dataset consisted of 47 malignant and 46 benign biopsy-confirmed nodules from 93 patients (74 females, 19 males; mean age 55.9±15.4 years). For each nodule, its principal transverse and longitudinal projection US images were included. Blinded to the diagnosis, an expert radiologist reviewed the images, outlined each nodule, and recorded descriptors of visual characteristics of nodules using the recently established ACR TIRADS consensus lexicon[8]. Viewing, outlining and annotation of the images was performed in the electronic Physician Annotation Device (ePAD)[14] (Figure 1). The images were preprocessed to reduce the US-specific artifacts such as speckle using nonlocal means based speckle filtering[15]. The advantage of this filtering method over traditional smoothing is its ability to preserve the edges, which is crucial for reliable extraction of computational features.

Semantic features for radiologist-based malignancy risk estimation

We derive a set of ten semantic features directly from the radiologist's annotations (Table 1). As reference malignancy risk estimators, we use existing expert annotation-based TIRADS classifiers. Using the ACR TIRADS lexicon, we reformulate and implement five such TIRADS classifiers as described in the works of Park et al.[9], Kwak et al. in 2011[10], Kwak et al. in 2013[11], Zayadeen et al.[12], and Russ et al.[13]

In their work, Park et al. [9] used 12 US features to assess the risk of malignancy using linear regression, and define six TIRADS categories based on the estimated risk. Kwak et al. in 2011[10] distinguished eight US features suspicious of malignancy, which were incorporated into a logistic regression-based malignancy risk quantification and the count-based classification system. Kwak et al. in 2013[11] used five US features of malignancy to build their weighted sum based TIRADS scoring system. Zayadeen et al.[12] proposed a scoring system based on the defined major and minor malignant features, and also benign features. Russ et al.[13] describes French TIRADS, a five-tier scoring system based on five malignant and six benign features. ACR TIRADS lexicon provides terms sufficient to describe these major US features of thyroid nodules and thus all five TIRADS classifiers were implemented using our expert annotations.

Table 1 The semantic features of thyroid nodules derived from consensus ACR TIRADS lexicon.

Semantic feature	Possible values of radiologist annotation
Composition	Solid Predominantly solid Predominantly cystic Cystic Spongiform
Echogenicity	Hyperechoic Isoechoic Hypoechoic Very hypoechoic
Shape	Taller-than-wide Wider-than-tall
Border	Smooth Irregular Lobulated Ill-defined
Halo	Present Absent
Extrathyroidal extension	Present Absent
Punctate echogenic foci	Present Absent
Macrocalcifications	Present Absent
Peripheral calcifications	Present Absent
Comet-tail artifacts	Present Absent

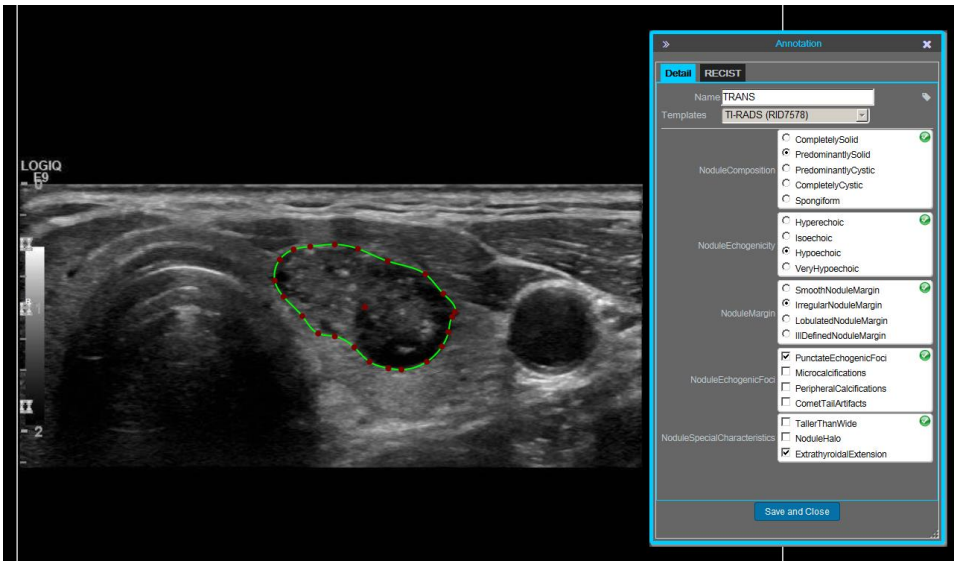

Figure 1 Example of annotation of a thyroid nodule on a transverse US image using the ePAD image viewing and annotation tool.

Computational features for automated malignancy risk estimation

To describe appearance of nodules identified by the radiologist in a quantitative manner, we selected a rich set of computational features that encode the echogenicity, texture, margin and shape properties of thyroid nodules. *Nodule echogenicity* was expressed by nodule intensity features, i.e., intensity histogram and intensity statistics. *Nodule composition and presence of echogenic foci* was expressed by texture and intensity features: rotation-invariant local binary pattern, Gabor filter bank, gray level co-occurrence matrix based features, and intensity difference between the nodule and its surrounding tissue. *Nodule margin* was computationally characterized via edge sharpness and local area integral invariant descriptor statistics at multiple spatial scales. *Nodule shape* had the following computational counterparts: ellipsoid fit, compactness, and roughness features. *Nodule halo and peripheral calcifications* were quantified by local intensity features, specifically by intensity histogram over the nodule margin. Although some of the semantic features, such as extrathyroidal extension, cannot be characterized easily from nodule outline alone, the edge sharpness features can be used as a surrogate, since the edges are expected to be blurred at the boundary portion where the thyroid capsule is invaded by the nodule[16]. Overall, 480 computational features were collected.

Classification

We formulate the automated malignancy risk estimation as a posterior probability of malignancy learned by a machine learning classifier on an annotated training dataset. For both radiologist-annotated semantic features and computer-derived computational features, we trained an RF classifier to predict whether the nodule is benign or malignant. Classification by RFs consists of creating an ensemble of decision tree (DT) classifiers, each learned on a subsample of the original training dataset, and predicting the outcome by averaging responses across the DTs. The subsamples are drawn with replacement and also consist of randomly selected features. The RFs are characterized by the number of grown trees and the depth of the trees (equivalent to the number of leaves) in the DTs.

We chose RF because it generally has superior performance in medical image analysis[17] and bioinformatics research[18], and it also has several additional desirable properties: 1) RFs can handle situations when the number of features are considerably higher than the number of data samples and do not require explicit dimensionality reduction as some other methods, 2) they can work with both categorical and continuous features and their combinations, and 3) they provide a probabilistic output. Therefore, in formulation of RF classification as discriminating benign vs malignant nodules, we will use the malignancy probability as the risk estimate.

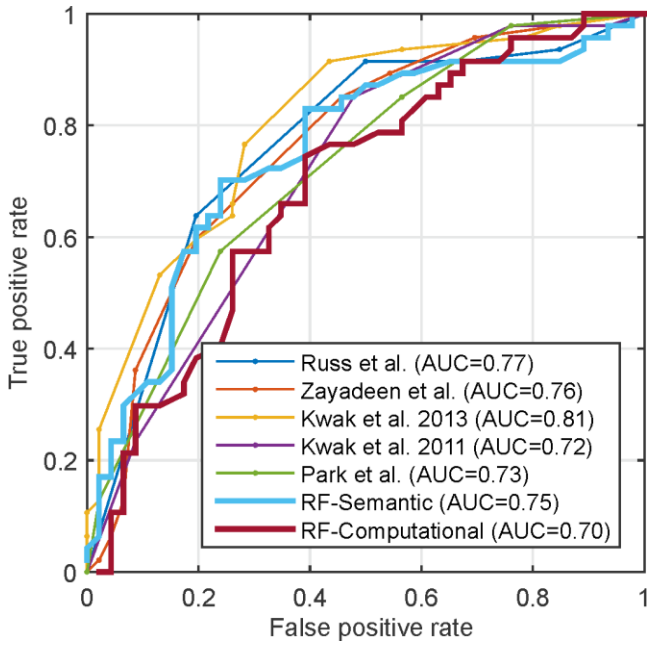

Figure 2 Performance of the proposed computational framework and five radiologist-based TIRADS scoring systems demonstrated using ROC curves.

Evaluation

We used five expert scoring systems, or TIRADS classifiers, as reference malignancy risk estimators. The annotations provided by the radiologist were transformed into a mineable feature set consisting of binary and categorical predictors. Overall, 10 semantic and 480 computational features were used to train RF classifiers (RF-Semantic and RF-Computational). Leave-one-out hyperparameter tuning and model fitting was performed. Within each fold, the minimal number of the leaves in the decision trees were selected by searching among values {2,5,10,20}, while the number of trees in the forest was fixed to value 50 across all folds. The RFs were trained on the biopsy-confirmed diagnoses of the nodules in two-label formulation, i.e., benign vs malignant.

We measured the predictive performance of the classifiers using receiver operating characteristic (ROC) curves built on the malignancy scores as predictions and nodule malignancy indicator as reference labels. As a quantitative evaluation metric, we used the corresponding values of area under the ROC curve (AUC), which is an estimate of the probability that, for the randomly selected malignant nodule and randomly selected benign nodule, their scores will be in the correct order, i.e. the malignant nodule will receive a score higher than the benign nodule.

Results

The performance summary of the five TIRADS and the two RF classifiers is demonstrated in Figure 2 using ROC curves. For TIRADS classifiers, performance in terms of AUC was AUC=0.73 for Park et al.[9], AUC=0.72 for Kwak et al. 2011[10], AUC=0.81 for Kwak et al. 2013[11], AUC=0.76 for Zayadeen et al.[12], and AUC=0.77 for Russ et al.[13] The proposed computational framework (RF-Computational) achieved AUC=0.70 and its semantic counterpart (RF-Semantic) provided AUC=0.75 (Figure 2). As expected, the classifiers that rely on radiologist annotation performed better than RF based on computational features alone, however, the performance was still comparable, which can also be noticed in Figure 2 as the ROC curves are at a close proximity to each other.

To study the clinical relevance of the methods, we analyzed their behavior at the highest true positive rate, or sensitivity. We recorded the number of correctly identified benign nodules, i.e., true negative count, in the absence or minimal number of misidentified malignant nodules, i.e. false negative count close to zero. An ideal classifier would correctly identify all benign nodules without missing any malignant nodules, thus resulting in the number of true negatives equal to the number of benign nodules while having false negative count as zero. The results on our dataset for all the implemented classifiers are summarized in Table 2. Although at non-zero missed malignancies (FN=1,2 or above) the TIRADS classifiers achieved higher number for spared benignities, implementation of such rules in practice

Table 2 Number of spared benign nodules (true negative count, TN) identified by the classifiers at a given count of missed malignancies (false negative count, FN) defined when available by setting accordingly the malignancy threshold value.

Classifier	Missed malignancies, FN			Next threshold	
	FN=0	**FN=1**	**FN=2**	**FN**	**TN**
Russ et al.	*0*	n/a*	n/a	3	7
Zayadeen et al.	*0*	7	14	5	21
Kwak et al. 2013	*0*	7	10	3	20
Kwak et al. 2011	*0*	11	n/a	7	24
Park et al.	*0*	11	n/a	7	20
RF-Semantic	*2*	3	5	3	5
RF-Computational	*5*	5	11	3	11

* no threshold provided the corresponding number of false negatives
Italics highlight the number of biopsies that could be avoided at no cost to other patients' health.

would result in a decrease of benign nodule biopsies at the cost of misdiagnosis of patients with cancer (FN>0). Only the RF classifiers were able to provide a positive number of true negatives with no false negatives, which is the number of biopsies that could be avoided at no cost to patients with malignancies.

Discussion

In this work, we presented a computational framework for evaluating thyroid nodules (benign vs. malignant) on US images, with performance comparable to that of expert annotation-based classification systems. The malignancy estimation performance with respect to the biopsy-confirmed diagnosis for our framework was 0.70 in terms of AUC, which was lower but comparable to performance of the classifier based on expert-annotated semantic features (AUC 0.75) and to values obtained by the expert annotation-based TIRADS classifiers (AUC ranged from 0.72 to 0.81). Comparability in terms of classification performance shows potential of such computational features to enhance or perhaps even replace annotation by radiologists, which is laborious, expensive and prone to intra- and interrater variability.

The clinical relevance of the classifiers as a potential pre-biopsy malignancy risk estimator was further analyzed in terms of the number of unnecessary biopsies that could have been avoided, i.e., nodules correctly identified as benign, at the minimal levels of misidentified malignancies. In our experiments, the proposed computational framework achieved the highest number of correctly identified benign nodules (5 out of 46) in the absence of missed malignancies (values in italics in Table 2). Although the TIRADS systems we compared our system against could identify more benign cases, they did so with more misidentified malignancies. Another disadvantage of TIRADS classifiers in this respect is the limited control over the threshold for benign nodule selection. This is due to the oversimplified rules that lay foundation to such qualitative classification systems and resulted in small number of possible thresholds, e.g. five nodule malignancy scores in the work by Park et al.[9] As can be seen in Table 2, only the RF classifiers provided thresholds for all the analyzed missed malignancy (false negative) counts.

Although some prior work has been done in thyroid cancer diagnosis using computational features derived from ultrasound images[19], the studies were limited to considerably smaller datasets (20 patients). More recently, Wu et al.[20] have shown that their framework of radiologist-annotated semantic features and machine learning classifiers can provide results comparable to that of expert malignancy scoring. However, to the best of our knowledge, our work is the first to demonstrate a framework of computational features and a machine learning classifier having a performance comparable to expert classification. Our computational framework could serve as a basis to develop computer-aided diagnosis tools that ultimately could help reduce the high number of unnecessary biopsies in patients with benign nodules.

Thyroid cancer diagnosis currently relies on cytopathological analysis, which is invasive and carries risks of bleeding, infection and potential damage to adjacent structures. Identifying patients who have low risk for cancer is important to avoid such unnecessary health risks and decrease societal healthcare costs. Recent attempts to address the issue by introducing more reliable pre-biopsy cancer risk estimation using standard-of-care US images are based on radiologist-

annotated standard descriptors, which are laborious to collect and subject to inter- and interrater variability[21]. Ultrasound is invaluable to noninvasive characterization of thyroid nodules, however, current guidelines are limited to recording a limited number of US feature of the nodules[8]. By computing a rich set of features from these images, our proposed framework maximizes the use of the information available in the images. The use of the computer-aided diagnosis tools that are based on frameworks similar to ours could improve the management of thyroid nodules and result in decreasing the number of unnecessary biopsy or surgical risks, while more appropriately directing care in patients who actually need more invasive management.

A limitation of this study was the use of only two principal projections, transverse and longitudinal, for extraction and analysis of semantic and computational features. Although the selection of these projections is standard in practice to evaluate nodule properties such as shape and size, the use of multiple image frames from the ultrasound exams could better other US features and will be analyzed in future work. Obtaining consistent US features is also challenged by the variability in positioning the transducer by the operator; an extended study with multiple operators to study inter- and intra-rater variability can give more insight on robustness of certain features and limitations of others. An additional limitation of our work was the use of a single radiologist for determining the semantic features, though that radiologist is an academic expert. Future studies assessing the impact of inter-reader variation in assessments among experts on our results could be helpful.

Conclusion

In this paper, we presented a computational framework for estimating the risk of cancer in thyroid nodules by analysis of US images. The framework computes a rich set of computation features in the US images and estimates the malignancy probability using random forest classifier. Performance of the framework in terms of AUC was lower but comparable to that of five baseline expert radiologist annotation-based TIRADS classifiers. Given that our framework does not require qualitative assessments from the expert radiologist, it could reduce the effort required to evaluate these images and reduce variation in practice. In addition, the number of biopsies that a classifier could help to avoid without missing any malignancies was the highest for our framework. Our results confirm the feasibility of computer-aided diagnosis systems for thyroid cancer risk estimation. Such systems could provide second-opinion malignancy risk estimation for clinicians and ultimately help decrease the number of unnecessary biopsies and surgeries.

Acknowledgements

This research was supported in part by grants from GE Medical Systems (Blue Sky Initiative at Stanford University) and the National Cancer Institute, National Institutes of Health, U01CA142555, 1U01CA190214, and 1U01CA187947.

References

1. Vaccarella S, Franceschi S, Bray F, Wild CP, Plummer M, Dal Maso L. Worldwide Thyroid-Cancer Epidemic? The Increasing Impact of Overdiagnosis. The New England Journal of Medicine. 2016 Aug 18;375(7):614–7.

2. Tan GH, Gharib H. Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. Annals of Internal Medicine. 1997 Feb 1;126(3):226–31.

3. Frates MC, Benson CB, Charboneau JW, Cibas ES, Clark OH, Coleman BG, et al. Management of Thyroid Nodules Detected at US: Society of Radiologists in Ultrasound Consensus Conference Statement. Radiology. 2005 Dec 1;237(3):794–800.

4. Howlader N, Noone A, Krapcho M, Miller D, Bishop K, Altekruse S, et al. SEER Cancer Statistics Review, 1975-2013, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2013/, based on November 2015 SEER data submission, posted to the SEER web site, April 2016.

5. Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, et al. Cancer statistics, 2005. CA: A Cancer Journal for Clinicians. 2005 Feb;55(1):10–30.

6. 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. 2015 Oct 14;26(1):1–133.

7. Yoon JH, Lee HS, Kim E-K, Moon HJ, Kwak JY. Malignancy Risk Stratification of Thyroid Nodules: Comparison between the Thyroid Imaging Reporting and Data System and the 2014 American Thyroid Association Management Guidelines. Radiology. 2015 Sep 8;278(3):917–24.

8. Grant EG, Tessler FN, Hoang JK, Langer JE, Beland MD, Berland LL, et al. Thyroid Ultrasound Reporting Lexicon: White Paper of the ACR Thyroid Imaging, Reporting and Data System (TIRADS) Committee. Journal of the American College of Radiology. 2015 Dec 1;12(12):1272–9.

9. Park J-Y, Lee HJ, Jang HW, Kim HK, Yi JH, Lee W, et al. A proposal for a thyroid imaging reporting and data system for ultrasound features of thyroid carcinoma. Thyroid. 2009 Nov;19(11):1257–64.

10. Kwak JY, Han KH, Yoon JH, Moon HJ, Son EJ, Park SH, et al. Thyroid Imaging Reporting and Data System for US Features of Nodules: A Step in Establishing Better Stratification of Cancer Risk. Radiology. 2011 Sep 1;260(3):892–9.

11. Kwak JY, Jung I, Baek JH, Baek SM, Choi N, Choi YJ, et al. Image reporting and characterization system for ultrasound features of thyroid nodules: multicentric Korean retrospective study. Korean Journal of Radiology. 2013 Feb;14(1):110–7.

12. Zayadeen AR, Abu-Yousef M, Berbaum K. Retrospective Evaluation of Ultrasound Features of Thyroid Nodules to Assess Malignancy Risk: A Step Toward TIRADS. AJR American Journal of Roentgenology. 2016 Jun 28;1–10.

13. Russ G, Royer B, Bigorgne C, Rouxel A, Bienvenu-Perrard M, Leenhardt L. Prospective evaluation of thyroid imaging reporting and data system on 4550 nodules with and without elastography. European Journal of Endocrinology. 2013 May 1;168(5):649–55.

14. Rubin DL, Willrett D, O'Connor MJ, Hage C, Kurtz C, Moreira DA. Automated Tracking of Quantitative Assessments of Tumor Burden in Clinical Trials. Translational Oncology. 2014 Feb 1;7(1):23–35.

15. Coupe P, Hellier P, Kervrann C, Barillot C. Nonlocal Means-Based Speckle Filtering for Ultrasound Images. IEEE Transactions on Image Processing. 2009 Oct;18(10):2221–9.

16. Kamaya A, Tahvildari AM, Patel BN, Willmann JK, Jeffrey RB, Desser TS. Sonographic Detection of Extracapsular Extension in Papillary Thyroid Cancer. Journal of Ultrasound in Medicine. 2015 Dec;34(12):2225–30.

17. Criminisi A, Shotton J. Decision Forests for Computer Vision and Medical Image Analysis. Springer; 2013. (Advances in Computer Vision and Pattern Recognition).

18. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine Learning methods for Quantitative Radiomic Biomarkers. Scientific Reports. 2015 Aug 17;5:13087.

19. Acharya UR, Vinitha Sree S, Krishnan MMR, Molinari F, Garberoglio R, Suri JS. Non-invasive automated 3D thyroid lesion classification in ultrasound: a class of ThyroScan^{TM} systems. Ultrasonics. 2012 Apr;52(4):508–20.

20. 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. American Journal of Roentgenology. 2016 Jun 24;207(4):859–64.

21. Liu YI, Kamaya A, Desser TS, Rubin DL. A Bayesian Network for Differentiating Benign From Malignant Thyroid Nodules Using Sonographic and Demographic Features. American Journal of Roentgenology. 2011 May 1;196(5):W598–605.