Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Research article

Comparison between linear and nonlinear machine-learning algorithms for the classification of thyroid nodules

Fu-sheng Ouyang[a,1], Bao-liang Guo[a,1], Li-zhu Ouyang[b,1], Zi-wei Liu[a], Shao-jia Lin[a], Wei Meng[a], Xi-yi Huang[c], Hai-xiong Chen[a], Qiu-gen Hu[a,*], Shao-ming Yang[a,*]

[a] *Department of Radiology, Shunde Hospital of Southern Medical University, Foshan, Guangdong, PR China*
[b] *Department of Ultrasound, Shunde Hospital, Southern Medical University, Foshan, Guangdong, PR China*
[c] *Department of Laboratory, Lecong Hospital of Shunde, Foshan, Guangdong, PR China*

ARTICLE INFO

Keywords:
Thyroid nodule
Ultrasonography
Diagnosis
Machine learning
Area under the curve

ABSTRACT

Background: A key challenge in thyroid carcinoma is preoperatively diagnosing malignant thyroid nodules. The purpose of this study was to compare the classification performance of linear and nonlinear machine-learning algorithms for the evaluation of thyroid nodules using pathological reports as reference standard.

Methods: Ethical approval was obtained for this retrospective analysis, and the informed consent requirement was waived. A total of 1179 thyroid nodules (training cohort, n = 700; validation cohort, n = 479) were confirmed by pathological reports or fine-needle aspiration (FNA) biopsy. The following ultrasonography (US) featu res were measured for each nodule: size (maximum diameter), margins, shape, aspect ratio, capsule, hypoechoic halo, composition, echogenicity, calcification pattern, vascularity, and cervical lymph node status. We analyzed five nonlinear and three linear machine-learning algorithms. The diagnostic performance of each algorithm was compared by using the area under the curve (AUC) of the receiver operating characteristic curve. We repeated this process 1000 times to obtain the mean AUC and 95% confidence interval (CI).

Results: Overall, nonlinear machine-learning algorithms demonstrated similar AUCs compared with linear algorithms. The Random Forest and Kernel Support Vector Machines algorithms achieved slightly greater AUCs in the validation cohort (0.954, 95% CI: 0.939–0.969; 0.954 95%CI: 0.939–0.969, respectively) than other algorithms.

Conclusions: Overall, nonlinear machine-learning algorithms share similar performance compared with linear algorithms for the evaluation the malignancy risk of thyroid nodules.

1. Introduction

Incidences of thyroid nodules and thyroid cancer are increasing worldwide, largely due to enhanced diagnostic practices [1]. Widespread use of ultrasonography (US) has exponentially increased thyroid nodule detection to about 20.0%–67.0% [2]. China has, by far, the largest population in the world [3]; thus, the burden of clinical management of thyroid nodules and thyroid cancer is enormous [3]. A recent large community-based study revealed a thyroid nodule prevalence of 49% in Chinese adults, when using conventional US examination. Nevertheless, only 5% of thyroid nodules are malignant [4,5]. Therefore, it is critical to differentiate malignant from benign thyroid nodules, and to avoid unnecessary fine-needle aspiration (FNA) biopsy and overtreatment, such as surgery. The early detection and distinction between benign and malignant thyroid nodules is particularly important to guide clinical treatment and choose operative methods.

Conventional US is the most commonly used method for evaluating thyroid nodules [6]. US imaging characteristics associated with malignant nodules include micro-calcifications, hypoechogenicity, microlobulated or irregular margin, taller-than-wide shape, rich vascularity on color Doppler, and presence of suspicious cervical lymph nodes [5,7]. However, the complex structures of the nodules lead to complex US images and some crossed signs between benign and malignant

Abbreviations: US, ultrasonography; Lasso, east absolute shrinkage and selection operator; RF, random forest; k-SVM, Kernel Support Vector Machines; Nnet, neural network; KNN, K-nearest neighborhood; NB, naive bayes; EN, elastic net; AUC, area under the curve; CI, confidence interval; FNA, fine-needle aspiration

[*] Corresponding authors at: No.1, Penglai Road, Daliang District, Shunde, Foshan, Guangdong, PR China.

E-mail addresses: qiugenhu@126.com (H. Qiu-gen), ysmsin@aliyun.com (S.-m. Yang).

[1] Fu-sheng Ouyang, Bao-liang Guo and Li-zhu Ouyang contributed equally to this work.

https://doi.org/10.1016/j.ejrad.2019.02.029
Received 15 July 2018; Received in revised form 13 February 2019; Accepted 21 February 2019

nodules [8]. Thus, operator-dependent US cannot satisfactorily distinguish malignant and benign nodules. It has been shown that machine-learning algorithms provide significantly better prediction than human experts [9]. Machine-learning algorithms can be divided into linear and nonlinear according to whether they need making strong assumptions about the form of the mapping function or not. Linear methods assume that features affect the final outcome only through a linear pattern. In contrast, nonlinear methods are more general and their performance is less dependent on the model assumption. Some nonlinear approaches such as extreme machine learning, convolutional neural networks, and artificial neural networks to thyroid nodule diagnosis and showed promising results [10,11]. However, that previous work trained models with relatively small sample and had not systematically explore which classifiers works better.

Therefore, the aim of this study was to assess and compare linear and nonlinear machine-learning algorithms for the evaluation of malignancy risk of thyroid nodules, and to identify the better algorithms for thyroid nodules.

2. Materials and methods

2.1. Patients and thyroid nodules

Ethical approval was obtained for this retrospective analysis, and the informed consent requirement was waived. The training and validation cohorts of this study were obtained from an evaluation of the medical database from January 2011 to November 2016, in order to identify patients with confirmed benign or malignant thyroid nodules who underwent conventional US examination. The inclusion criteria for nodules in this study were as follows: 1) underwent FNA cytology and/or surgery histology within 1 month after examinations; 2) had complete medical information and no prior surgical treatment performed on the nodules. Exclusion criteria were: 1) inadequate cytologic results and lack of surgical histology (n = 123); 2) indeterminate cytologic results and lack of surgical histology (n = 34); 3) diagnosis as “suspicious for papillary thyroid carcinoma” in cytologic examination and lack of surgical histology (n = 11).

2.2. Real-time grayscale US assessment of thyroid nodules

We used Acuson Sequoia 512 (Siemens Medical Solutions) and 128XP sonographic scanners (Siemens Medical Solutions) equipped with 10–12 MHz linear probes. US scans of both thyroid lobes and isthmus were obtained in both transverse and longitudinal planes. Longitudinal and transverse images of the thyroid are obtained according to the ACR accreditation standards, and images were recorded on picture archiving and communication system (PACS) workstations (GE Centricity; GE Healthcare). Sonographers conducted the initial image acquisition. Grayscale US features of each thyroid nodule were measured by two independent sonographers with 15 and 17 years of experience in thyroid US imaging, respectively, who were blinded to the pathology. Conflicts between the readers were resolved by consensus with a third reader. The following features were measured for each nodule: size, margins, shape, aspect ratio, capsule, hypoechoic halo, internal composition, echogenicity, calcification pattern, vascularity on color Doppler, and cervical lymph node status. The size of the nodule was its widest diameter. Margins were classified as well-defined or poorly defined. Shape was regular or irregular. Aspect ratio (height divided by width on transverse views) was classified as < 1 or ≥1. Thyroid capsule was categorized as complete or incomplete; in the latter case, tumoral tissue extended beyond the contours of the thyroid gland and invaded into adjacent structures. Hypoechoic halo was categorized as usual or unusual; in the latter case, the lesion was surrounded by an irregular or thick hypoechoic halo. Internal composition was classified as cystic, solid, or mixed (solid portion ≤50%). Echogenicity was defined as absence, hyperechogenicity, isoechogenicity, or hypoechogenicity. Calcification pattern was categorized into no calcification, microcalcification, macrocalcification, or both. Vascularity was assessed as the presence or absence of blood flow on color Doppler. Cervical lymph node status was assigned as normal or abnormal, based on US findings of calcification or cystic changes. Cervical lymph nodes were considered enlarged when the aspect ratio was > 0.5 without calcifications or cystic changes.

2.3. Linear machine-learning algorithms

For the linear machine-learning algorithms, we analyzed the Ridge, Lasso-penalty, and Elastic Net (EN) algorithms.

2.3.1. Ridge regression

The Ridge regression method is the most widely applied solution for addressing problems of multicollinearity, also showing better performance than other approaches when the sample size is small [12,13]. Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares,

$$\hat{\beta}^{ridge} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{1}$$

Here λ ≥ 0 is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. The coefficients are shrunk toward zero (and each other). Ridge regression has the ability to control the smoothness of a fitted model, thus avoiding overfitting, which can lead to a better prediction performance [14].

2.3.2. Lasso-penalty

Lasso-penalty is a suitable method for the regression of high-dimensional data, it tends to produce some coefficients that are exactly 0 and hence gives interpretable

models and avoids model overfitting [15]. The lasso is a shrinkage method like ridge, with subtle but important differences. The Lasso estimate is defined by

$$\overset{\wedge lasso}{\beta} = \underset{\beta}{\text{argmin}} \sum\nolimits_{i=1}^{N} \left| \mathrm{y_i} - \beta_0 - \sum\nolimits_{j=1}^{p} \mathrm{x_{ij}}\beta_j \right|^2 \text{subject to} \sum\nolimits_{j=1}^{p} |\beta_j| £\mathrm{t} \tag{2}$$

Just as in ridge regression, we can re-parametrize the constant by standardizing the predictors; the solution for $\hat{\beta}_0$ is $\bar{y}$, and thereafter we fit a model without an intercept. In the signal processing literature, the lasso is also known as basis pursuit. We can also write the lasso problem in the equivalent Lagrangian form

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3}$$

Notice the similarity to the ridge regression problem: the L_2 ridge penalty $\sum_i^p \beta_j^2$ is replaced by the L1 lasso penalty $\sum_i^p |\beta_j|$. This latter constraint makes the solutions nonlinear in the yi, and there is no closed form expression as in ridge regression. Computing the lasso solution is a quadratic programming problem, although we see that efficient algorithms are available for computing the entire path of solutions as λ is varied, with the same computational cost as for ridge regression. Because of the nature of the constraint, making t sufficiently small will cause some of the coefficients to be exactly zero. Therefore the lasso does a kind of continuous subset selection. If t is chosen larger than $t_0 = \sum_1^p |\hat{\beta}_j|$ (where $\hat{\beta}_j = \hat{\beta}_j^{ls}$ the least squares estimates), then the lasso estimates are the $\hat{\beta}_j$’ s. On the other hand, for $t = \frac{t_0}{2}$ say, then the least squares coefficients are shrunk by about 50% on average.

2.3.3. EN

Similar to the lasso, the EN simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables [16]. It is like a stretchable fishing net that retains 'all the big fish'. The EN is the combination of the Lasso-penalty and Ridge penalties. It overcomes several potential limitations of Lasso-penalty. which uses a penalty function based on

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \tag{4}$$

Use of this penalty function has several limitations. For example, in the "large p, small n" case (high-dimensional data with few examples), the lasso selects at most n variables before it saturates. Also if there is a group of highly correlated variables, then the lasso tends to select one variable from a group and ignore the others. To overcome these limitations, the EN adds a quadratic part to the penalty ($\|\beta\|^2$), which when used alone is ridge regression (known also as Tikhonov regularization). The estimates from the EN method are defined by

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1) \tag{5}$$

The quadratic penalty term makes the loss function strictly convex, and it therefore has a unique minimum. The EN method includes the lasso and ridge regression: in other words, each of them is a special case where $\lambda_1 = \lambda,\ \lambda_2 = 0\, or\, \lambda_1 = 0,\ \lambda_2 = \lambda$. Meanwhile, the naive version of EN method finds an estimator in a two-stage procedure: first for each fixed λ_2 it finds the ridge regression coefficients, and then does a lasso type shrinkage. This kind of estimation incurs a double amount of shrinkage, which leads to increased bias and poor predictions. To improve the prediction performance, the authors rescale the coefficients of the naive version of elastic net by multiplying the estimated coefficients by $(1+\lambda_2)$.

2.4. Nonlinear machine-learning algorithms

For the nonlinear algorithms, we analyzed the random forest (RF), kernel-Support Vector Machines (k-SVM), Neural Network (Nnet), kernel nearest neighborhood (k-NN), and Naïve Bayes (NB).

2.4.1. RF

RF has the ability to avoid the high variance or bias in prediction and is very simple to be implemented efficiently and accurately. It also has an effective method for estimating missing data and maintains accuracy when a large proportion of the data is missing. RFs are designed to further improve the accuracy of models by building multiple trees i.e. a Forest [17]. To make a prediction for a new observation, each tree in the forest gives a classification (a vote) on the outcome and the forest choose the classification having the majority votes (over all trees in the forest). To build a RF, each constituent tree is forced to split on only a random subset of the available independent variables from a bootstrapped sample of the data. For example, the training data for each tree is selected randomly with replacement-since each tree uses different independent variables and different training data, we generate a Forest of different trees.

RF is a modification of bagging that builds and averages many trees to obtain an approximately unbiased models to reduce the variance. Precisely, the RF will grow B trees to the bootstrapped data and repeating some statistical procedure to finally return the ensemble of the grown trees. When a new observation x is coming, let the $\hat{D}^b{}_{RF}(X)$ be the class prediction of the b-th tree, the

$$\hat{\mathrm{D}}_{\mathrm{RF}}(\mathrm{x}) = \text{majority vote}\{\widehat{\mathrm{D}^{\mathrm{b}}}_{\mathrm{RF}}(\mathrm{X})\}_{\mathrm{b}=1}^{\mathrm{B}} \tag{6}$$

2.4.2. k-SVM

SVM is of robustness for high-dimensional data and has good generalization capabilities, however, it is more difficult to interpret the generated model and has a certain sensitivity to appropriate parameter adjustments [18]. SVM can be formulated in the context of binary classification as the model that finds the decision boundary which maximizes the margin between two data classes. The Gaussian kernel is applied in the kernel-SVM, which is a very powerful tool in classification and has great performance in different kinds of real-life applications [14]. Precisely, considering the two class binary classification problem, the SVM aims to find a hyperplane which maximizes the margin between two classes. Generally, the classifier is assumed to be contained in a reproducing kernel Hilbert space (RKHS) induced by some kernel function K, denoted as HK. More importantly, the classifiers can be learnt from high dimensional features spaces and represent by a form of finite linear combination of the kernel functions. It is well known that any continuous function can be well approximated by the function in the RKHS if the kernel K is universal, for example, the Gaussian kernel. The classification problem of SVM can be formulated to solve the following optimization problem

$$\min_{f\in H_k} \frac{1}{n}\sum_{i=1}^{n} (1 - y_i f(x_i)) + \lambda \|f\|_{H_k}^2 \tag{7}$$

Such a problem can be efficiently solved by transforming the objective function to its dual problem and applying the KKT condition. By the representer theorem, the solution of such a optimization problem must satisfy that

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i K(x_i, x) \tag{8}$$

Where $\hat{\alpha}_i$ is the representer coefficients.

2.4.3. Nnet

Nnet is a very popular classification tool which has the ability to capture the potential nonlinear relationships between the variables, even the interaction terms [14]. Nnet is also called the single hidden layer back-propagation network or the single layer networks. Actually, a Nnet is a two stage classification method.

2.4.4. k-NN

k-NN is one of the easiest and simplest methods to predict classification in pattern recognition [18]. It is a instance-based learning method, where the classification function is approximated by the local data [14]. k-NN uses the training samples to predict a new sample by a majority vote on the outcome of the k-nearest points to the new sample. Given the sample $\{(y_i, x_i)\}_{i=1}^{n}$, the k-nearest neighbor uses those observations in the training set which are closest to x to predict y. In detail, by using the Euclidean distance as a metric to measure the closeness, the k-NN can be fitted that

$$\hat{y} = \frac{1}{k_{x_i}} \sum_{\hat{I}N_k(x)} y_i \tag{9}$$

where $Nk\,(x)$ is the neighborhood of x defined by the k closest points x_i in the training set.

2.4.5. NB

NB is one of the simplest probabilistic classifiers. NB is a probabilistic based method, which applies Bayes' theorem with strong (naive) independence assumptions between the variables. NB converges better than some discriminative models (e.g., logistic regression model), and thus requires fewer training sample to achieve an accurate classification results [14]. NB is a popularly used classification method which bases on Baye's theorem with the 'naive' assumption of independence between every pair of features. Give the pair $(y;\, x)$, where y = {-1, 1} is the label and $x = (x_1, \ldots, x_\mathrm{p})^\mathrm{T}$ denotes the variables. The Baye's theorem is formulated as

$$p(y|x) = \frac{P(y)P(x|y)}{P(x)} \tag{10}$$

With the naive independence assumption, the classifier is constructed:

$$\hat{y} = arg \max_{y\in\{-1,1\}} P(y) \prod_{i=1}^{p} P(x_i|y) \tag{11}$$

Then the label can be assigned by the probability model.

2.5. *Diagnostic performance of the linear and nonlinear machine-learning algorithms in the training and validation cohorts*

Using each of these algorithms, data were analyzed 1000 times after randomly dividing the entire sample of 1179 nodules into a training set (n = 700) and a validation set (n = 479). To quantify the diagnostic performance of the machine-learning algorithms, AUC was measured. Internal validation was performed by using the remaining data set after selection of the training set. The machine-learning algorithms formed in the training cohort were applied to all nodules of the validation cohort, with the total points calculated for each nodule. By outputting the probability of malignancy, our algorithms could assign the thyroid nodule(s) of a patient to a pattern within either Thyroid Imaging Reporting and Data Systems (TI-RADS) or American Thyroid Association (ATA) Management Guidelines. Using a fixed random number seed to generate a subset of the data for this validation (model versus radiologist) [19], we compared the diagnostic performance of the machine learning algorithms with visual inspection by blinded radiologists based on US features. The performance of radiologist's diagnosis was assessed by receiver operating characteristic curve analysis.

2.6. *Statistical analysis*

Statistical analysis was performed by using R version 3.2.3 (R Foundation for Statistical Computing, Vienna, Austria). The following R packages were used for classification algorithms: "randomForest" (RF), "e1071″ (k-SVM), "knn" (k-NN), "nnet" (Nnet), (Ridge-penalty), "LiblineaR" (Lasso-penalty), "e1071″ (NB), and "glmnet" (EN). We repeated the analyses 1000 times to obtain the mean AUC and 95% CI. We compared the AUCs of linear and nonlinear machine-learning algorithms. Inter- and intra-operator agreements were estimated by Cohen's Kappa. For interpretation, a Kappa coefficient > 0.7 was considered good agreement.

3. Results

3.1. *Basic characteristics of patients and nodules*

There were 1179 thyroid nodules (training cohort, n = 700; validation cohort, n = 479) in 1036 patients, including 501 (42.5%) benign and 678 (57.5%) malignant nodules. Among 1036 patients, 311 were men and 725 were women; the mean age (± standard deviation, SD) of the patients was 42.4 ± 12.8 years (range, 18–85 years). A total of 87, 263, 65, and 86 benign nodules were classified as TI-RADS II, III, IV, and V, respectively; 5, 81, 54, and 538 malignant nodules were categorized as TI-RADS II, III, IV, and V, respectively. Conventional US features of these nodules are summarized in Table 1.Intra- or inter-operator agreement of conventional US features were between 0.76 and 0.98 (Table 2).

Table 1
Conventional US according to Malignancy and Benignity.

Feature	No. of Benign Nodules (n = 501)	No. of Malignant Nodules (n = 678)	*P* value*
Margins			< .001
well-defined (n = 564)	472 (94.2)	92 (13.6)	
ill-defined (n = 615)	29 (5.8)	586 (86.4)	
Shape			< .001
Regular (n = 532)	471 (94.0)	61 (9.0)	
Irregular (n = 647)	30 (6.0)	617 (91.0)	
Aspect Ratio			< .001
< 1 (n = 830)	489 (97.6)	341 (50.3)	
≥1(n = 349)	12 (2.4)	337 (49.7)	
Capsule			< .001#
Complete (n = 1023)	498 (99.4)	525 (77.4)	
Incomplete (n = 156)	3 (0.6)	153 (22.6)	
Hypoechoic halo			< .001
Usual (n = 1146)	493 (98.4)	653 (96.3)	
Unusual (n = 33)	8 (1.6)	25 (3.7)	
Internal composition			< .001#
Cystic (n = 31)	31 (6.2)	0	< .001
Solid (n = 1077)	422 (84.2)	655 (96.6)	< .001
Mixed (n = 71)	48 (9.6)	23 (3.4)	< .001
Echogenicity			< .001#
Absence (n = 31)	31 (6.2)	0	< .001
Hypoechogenicity (n = 722)	101 (20.2)	621 (91.6)	< .001
Isoechogenicity (n = 365)	308 (61.5)	57 (8.4)	< .001
Hyperechogenicity (n = 61)	61	0	
Calcification patterns			< .001#
No calcification (n = 605)	466 (93.0)	139 (20.5)	< .001
Microcalcification (n = 444)	3 (0.6)	441 (65.0)	< .001
Macrocalcification (n = 80)	31 (6.2)	49 (7.2)	0.483
Micro + macrocalcification (n = 50)	1 (0.2)	49 (7.2)	< .001
Cervical lymph node status			< .001#
Absent (n = 928)	486 (97.0)	442 (65.2)	< .001
Calcifications or cystic (n = 117)	0	117 (17.3)	< .001
Enlarged (n = 134)	15 (3.0)	119 (17.5)	< .001
Vascularity on color Doppler			< .001
Absence (n = 1036)	478 (95.4)	558 (82.3)	
Presence (n = 143)	23 (4.6)	120 (17.7)	

Note: unless otherwise indicated, data are numbers of nodules, and numbers in parentheses are percentages.

* P value were calculated by using generalized estimating equation analysis.

# Fisher exact test.

Table 2
Intra- and inter-operator conventional US feature measurement reproducibility: coefficients of variation.

Features	Intra-operator		Inter-operator	
	Operator 1	Operator 2	Measurement 1	Measurement 2
Size	0.91		0.95	
Echogenicity	0.79		0.89	
Margins	0.76		0.88	
Shape	0.78		0.89	
Aspect ratio	0.92		0.98	
Calcification patterns	0.85		0.92	
Capsule	0.87		0.96	
Internal composition	0.86		0.93	
Hypoechoic halo	0.82		0.89	
Cervical lymph node status	0.83		0.91	
Vascularity on color Doppler	0.81		0.88	

3.2. *Diagnostic performance of the linear and nonlinear machine-learning algorithms in the training and validation cohorts*

The diagnostic performance of all machine-learning models was based on 11 US imaging features. In the training set, mean AUCs for the linear machine-learning algorithms Ridge-penalty, Lasso-penalty, and EN were 0.937 (95% CI: 0.927-0.947), 0.936 (95% CI: 0.918-0.953), and 0.937 (95% CI: 0.927-0.947). Mean AUCs for the nonlinear machine-learning algorithms RF, k-SVM, Nnet, k-NN, and NB were 0.989 (95% CI: 0.975–1.000), 0.971 (95% CI: 0.963-0.979), 0.936 (95% CI: 0.902-0.969), 0.971 (95% CI: 0.956 to 0.986), and 0.942 (95%

Table 3
Diagnostic performance of linear and nonlinear machine learning algorithms and experienced radiologists for the evaluation of thyroid nodules.

	AUC (95%CI)	
	Training cohort	Validation cohort
Linear models		
Ridge-penalty	0.937 (95%CI: 0.927-0.947)	0.936 (95%CI: 0.920-0.952)
Lasso-penalty	0.936 (95%CI: 0.918-0.953)	0.935 (95%CI: 0.915-0.954)
EN	0.937 (95%CI: 0.927-0.947)	0.936 (95%CI: 0.920-0.952)
Nonlinear models		
RF	0.989 (95%CI: 0.975-1.000)	0.954 (95%CI: 0.940-0.968)
k-SVM	0.971 (95%CI: 0.963-0.979)	0.954 (95%CI: 0.939-0.969)
Nnet	0.936 (95%CI: 0.902-0.969)	0.928 (95%CI: 0.883-0.973)
k-NN	0.971 (95%CI: 0.956-0.986)	0.937 (95%CI: 0.919-0.955)
NB	0.942 (95%CI: 0.932-0.951)	0.940 (95%CI: 0.925-0.950)
Radiologist	—	0.830 (95%CI: 0.811-0.849)

Note: Lasso-least absolute shrinkage and selection operator; RF, random forest; k-SVM, kernel Support Vector Machines; Nnet-Neural Network; k-NN, k-nearest neighborhood; NB-naive bayes; EN, elastic net.

CI: 0.932-0.951). Thus, the RF algorithm outperformed other machine-learning algorithms in the training cohort (Table 3).

Application of the nonlinear machine-learning algorithms RF, k-SVM, Nnet, k-NN, and NB in the internal validation cohort yielded high mean AUCs: 0.954 (95% CI: 0.940-0.968), 0.954 (95% CI: 0.939-0.969), 0.928 (95% CI: 0.883-0.973), 0.937 (95% CI: 0.919-0.955), and 0.940 (95% CI: 0.925-0.950). Mean AUCs for Ridge-penalty, Lasso-penalty, and EN were 0.936 (95% CI: 0.920-0.952), 0.935 (95% CI: 0.915-0.954), and 0.936 (95% CI: 0.920-0.952). Thus, the RF and k-SVM algorithms outperform other algorithms in the validation cohort (Table 3). The AUCs of machine learning algorithms were larger than experienced radiologists without the assistance of machine learning (AUC, 0.830, 95%CI: 0.811-0.849).

3.3. Relative importance of each feature within the RF algorithm

The diagnostic performance of the RF algorithm was most dependent on the following five top-rank features, according to their mean decrease in Gini: calcification pattern (132.3), shape (62.6), cervical lymph node status (58.1), margins (50.7), and aspect ratio (46.7) (Fig. 1).

3.4. Examples of RF and k-SVM algorithms in use

For example, patient 1 had multiple cystic thyroid nodules. The largest nodule had

the following features: maximum diameter of 0.7 cm, well-defined margins, regular shape, aspect ratio < 1, complete capsule, no echo, absence of hypoechoic halo, no calcification, no color Doppler flow imaging (CDFI) signal, no enlarged cervical lymph node (Fig. 2a-b). The nodule was assigned a TI-RADS II. The malignancy risk calculated at 0% by RF and k-SVM algorithms. FNA biopsy confirmed it was a follicular cyst. Patient 2 had a solid nodule in the left leaf of the thyroid, with maximum diameter of 0.9 cm, well-defined margins, regular shape, aspect ratio < 1, hypoechogenicity, absence of hypoechoic halo, no calcification, litter CDFI signal and no enlarged cervical lymph node (Fig. 2c-d). The nodule was assigned a TI-RADS III. The malignancy risk calculated at 2% by RF and k-SVM. Pathology confirmed it was a nodular goiter. Patient 3 had a solid nodule in the right leaf of the thyroid, with maximum diameter of 0.7 cm, ill-defined margins, irregular shape, aspect ratio > 1, hypoechogenicity, absence of hypoechoic halo, microcalcification, little CDFI signal around the nodule and no enlarged cervical lymph node (Fig. 2e-f). The nodule was assigned a TI-RADS IV. The malignancy risk calculated at about 83% by RF and k-SVM. Pathology confirmed it was a papillary carcinoma. Patient 4 had a solid nodule in the right leaf of the thyroid, with maximum diameter of 0.9 cm, ill-defined margins, irregular shape, aspect ratio = 1, incomplete capsule, hypoechogenicity, microcalcification, rich CDFI signal, and an enlarged cervical lymph node (Fig. 2g-h). The nodule was assigned a TI-RADS V. The malignancy risk of could be calculated at about 90% by RF and k-SVM. Pathology confirmed it was a papillary carcinoma.

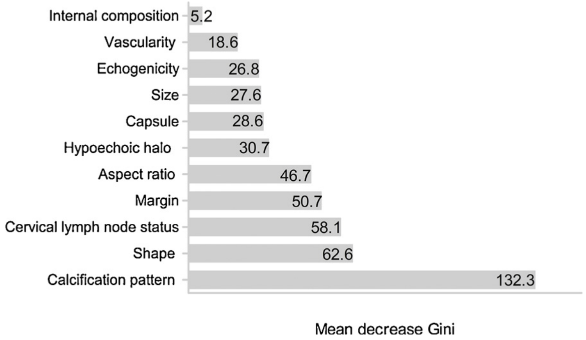

Fig. 1. Relative importance (mean decrease Gini) of each feature within the RF algorithm.

4. Discussion

We compared three linear and five nonlinear machine-learning algorithms for the evaluation of thyroid nodules. Overall, the nonlinear algorithms demonstrated similar performance with the linear algorithms. Our results showed that the nonlinear machine-learning algorithm RF and k-SVM had slightly better performance than other methods. This easy-to-use, repeatable, and low-cost machine-learning algorithm may facilitate the diagnosis of malignant nodules.

FNA has been widely accepted as the preferred detection technique for the assessment of non-toxic thyroid nodules; it aims primarily to triage patients, in order to identify those requiring surgery and then select appropriate surgical procedures [20]. However, thyroid FNA is a limited technique, mainly because of uncertain cytological features in differentiating benign from malignant follicular neoplasms, Hürthle cells, hyperplastic nodules, and follicular variants of papillary cancer [21,22]. Thyroid FNA results are non-diagnostic, suspicious, or indeterminate in 20%–30% of cases, and the risk of cancer varies from 5%–75.5% in the above-mentioned group [21]. Recommended clinical management for indeterminate aspirates is repeated FNA. However, repeated FNA exhibits a 38%–48% indeterminate rate. Additionally, diagnostic surgery may be performed; this eventually increases medical costs and patient suffering [22].

This present study reported common US features of thyroid nodules, including size, margins, shape, aspect ratio, capsule, hypoechoic halo, internal composition, echogenicity, calcification pattern, vascularity on color Doppler, and cervical lymph node status. A previous study showed that US characteristics, rather than nodule size, should be used to differentiate between nodules, which suggested that thyroid nodules of any size should be further examined for malignancy [23]. Our study also indicated the nodule size plays a very minor role in the prediction of malignancy. In the majority of cancer cases, a single nodule is present within a cluster. Most of nodules in the present study were single. No significant relationships were found between number of the nodules and malignancy [24]. Most researchers agree that higher numbers of nodules do not increase the risk of malignancy [25]. The previous studies including meta-analyses compared US and FNA results of thyroid nodules to US criteria for the diagnosis of malignancy, including micro-calcification, hypo-echogenicity, taller-than-wide shape, and ill-defined margins [26,27]. Within the machine-learning algorithms such as RF, calcification pattern, shape, cervical lymph node status, margins, and aspect ratio were the most strong predictors of malignancy, which was consistent with the previous studies. Unlike

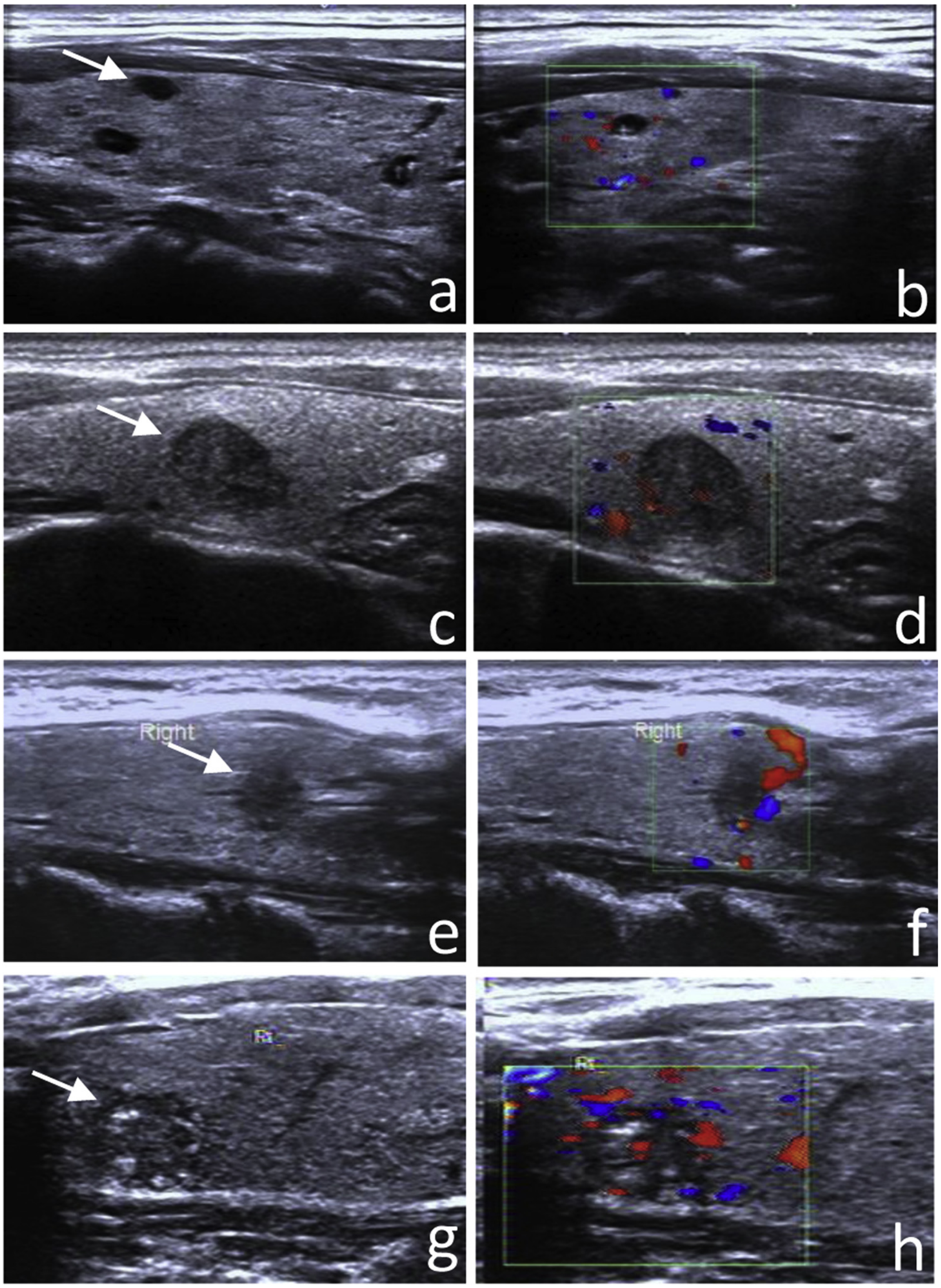

Fig. 2. US images of four representative cases. Case 1 (a–b): The nodule (white arrow) had diameter of 0.7 cm, well-defined margins, regular shape, aspect ratio < 1, complete capsule, no echo, no hypoechoic halo, no calcification, no CDFI signal, and no enlarged cervical lymph node. The malignancy risk calculated at 0% by RF. **Case 2 (c–d):** A solid nodule (white arrow) located in the left leaf, with diameter of 0.9 cm, well-defined margins, regular shape, aspect ratio < 1, hypoechogenicity, no hypoechoic halo, no calcification, litter CDFI signal and no enlarged cervical lymph node. The malignancy risk calculated at 2% by RF and k-SVM. **Case 3 (e–f).** A solid nodule (white arrow) located in the right leaf, with diameter of 0.7 cm, ill-defined margins, irregular shape, aspect ratio > 1, hypoechogenicity, no hypoechoic halo, microcalcification, little CDFI signal and no enlarged cervical lymph node. The malignancy risk calculated at about 83% by RF and k-SVM. **Case 4 (g–h):** A solid nodule (white arrow) located in the right leaf, with diameter of 0.9 cm, ill-defined margins, irregular shape, aspect ratio = 1, incomplete capsule, hypoechogenicity, microcalcification, rich CDFI signal, and an enlarged cervical lymph node. The malignancy risk of could be calculated at about 90% by RF and k-SVM.

other studies, we evaluated the cervical lymph nodes, which was a vital part of every thyroid sonographic examination with high specificity. In this study, its predictive value was next only to that of calcification pattern. A cystic lymph node ipsilateral to a thyroid nodule should be nearly pathognomonic of metastatic papillary thyroid carcinoma.

The present work possesses advantages over several studies that have attempted to differentiate malignant from benign thyroid nodules using computer-aided diagnosis (CAD) systems with machine learning on the basis of US features [28–32]. The CAD systems showed mixed results when compared with visual inspection by experienced radiologists. Choi YJ et al. found a similar sensitivity as the experienced radiologist (90.7% vs. 88.4%, p > 0.99), but a lower specificity and a lower AUC (specificity: 74.6% vs. 94.9%, p = 0.002; AUC: 0.83 vs. 0.92, p = 0.021) [28]. Gitto S et al. reported the radiologist showed a significantly higher sensitivity than the CAD system (78.6% vs. 21.4%; P = 0.008), while there was no statistical difference in specificity (66.7% vs. 81.3%; P = 0.065) [30]. Jeong EY et al. observed the sensitivity and accuracy of the CAD system did not differ significantly from those of the experienced radiologist (88.6% vs. 84.1%, p = 0.687; 86.0% vs. 91.0%, p = 0.267) while the specificity was significantly higher for the experienced radiologist (83.9% vs. 96.4%, p = 0.016) [31]. Also, Chang Y et al. concluded that the computed AUC were 0.986 and 0.979 for the proposed CAD system and visual inspection by radiologists, with no significantly difference [32]. In this study, the machine learning models had higher AUC than experienced radiologists, especially for RF and k-SVM methods (0.954 vs 0.830).

However, those CAD systems classify US features differently from the way radiologists do, whereas our models examine the same features as radiologists. Our machine-learning models examines a larger array of US features than previous CAD studies. In addition, this present study compared nonlinear and linear methods to identify the best model in dealing with the real thyroid data. Last but not least, the prediction process of our models was easier than CAD systems because the models required no preprocessing procedure of the images and extraction textural features from the images.

Despite these strong results, our study has a number of limitations. Firstly, we did not analyze additional US characteristics, such as composition or number of nodules, because they showed no significant predictive value, according to most previous studies. Secondly, we could not distinguish encapsulated follicular variant of papillary thyroid carcinoma from malignant thyroid nodules by US, though it has excellent prognosis and thus is described as a noninvasive follicular thyroid neoplasm with papillary-like nuclear features [33]. In addition, it was newly reported, so our pathology department had not differentiated it from malignant nodules. US images taken with different settings (gain, focus etc.) may cause variation in B-mode intensity values. However, to reduce the potential variation, US scanners were equipped with suitable conditions for thyroid US imaging (including probe depth, frame frequency, gain of B-mode, dynamic range, gain, and focus). Lastly, the diagnostic performance of our classifiers depends on the operator-reported imaging features.

In summary, we developed and validated multiple machine-learning algorithms for the prediction of malignant thyroid nodules. It identifies those nodules with an acceptably low risk of malignancy to avoid numerous FNAs. However, the predictive performance of these algorithms should be further investigated in prospective cohorts.

Conflict of interest

The authors have declared that no conflict of interest exists

Funding

This work was supported by the science and technology project of foshan (2017AB003623 and 2017AB003683).

References

[1] J.H. Yoon, H.S. Lee, E.K. Kim, H.J. Moon, J.Y. Kwak, Malignancy risk stratification of thyroid nodules: comparison between the thyroid imaging reporting and data system and the 2014 american thyroid association management guidelines, Radiology 278 (2016) 917–924.

[2] H. Gharib, E. Papini, R. Paschke, D.S. Duick, R. Valcavi, L. Hegedus, et al., American Association of Clinical Endocrinologists, Associazione Medici Endocrinologi, and European Thyroid Association Medical Guidelines for Clinical Practice for the Diagnosis and Management of Thyroid Nodules, Endocr. Pract. 16 (Suppl. 1) (2010) 1–43.

[3] M. Xing, Clinical utility of RAS mutations in thyroid cancer: a blurred picture now emerging clearer, BMC Med. 14 (2016) 12.

[4] K.D. Burman, L. Wartofsky, CLINICAL PRACTICE. Thyroid nodules, N. Engl. J. Med. 373 (2015) 2347–2356.

[5] B.R. Haugen, E.K. Alexander, K.C. Bible, G.M. Doherty, S.J. Mandel, Y.E. Nikiforov, et al., 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) (2015) 1–133.

[6] K. Kaur, N. Sonkhya, A.S. Bapna, P. Mital, A comparative study of fine needle aspiration cytology, ultrasonography and radionuclide scan in the management of solitary thyroid nodule: a prospective analysis of fifty cases, Indian J. Otolaryngol. Head Neck Surg. 54 (2002) 96–101.

[7] H.J. Tae, D.J. Lim, K.H. Baek, W.C. Park, Y.S. Lee, J.E. Choi, et al., Diagnostic value of ultrasonography to distinguish between benign and malignant lesions in the management of thyroid nodules, Thyroid 17 (2007) 461–466.

[8] J.H. Shin, J.H. Baek, J. Chung, et al., Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean society of thyroid radiology consensus statement and recommendations, Korean J. Radiol. 17 (3) (2016) 370–395.

[9] C. Oberije, G. Nalbantov, A. Dekker, L. Boersma, J. Borger, B. Reymen, et al., A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: a step toward individualized care and shared decision making, Radiother. Oncol. 112 (2014) 37–43.

[10] J. Xia, H. Chen, Q. Li, et al., Ultrasound-based differentiation of malignant and benign thyroid nodules: an extreme learning machine approach, Comput. Methods Programs Biomed. 147 (2017) 37–49.

[11] L.C. Zhu, Y.L. Ye, W.H. Luo, et al., A model to discriminate malignant from benign thyroid nodules using artificial neural network, PLoS One 8 (12) (2013) e82211.

[12] M. El-Dereny, N. Rashwan, Solving multicollinearity problem using ridge regression models, Int. J. Contemp. Math. Sci. 6 (2011) 585–600.

[13] A. Fitrianto, L.C. Yik, Performance of Ridge Regression Estimator Methods on small sample size by varying correlation coefficients: a simulation study, J. Math. Stat. 10 (2014) 25, https://doi.org/10.3844/jmssp.2014.25.29.

[14] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd edition, Springer, New York, 2009.

[15] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.

[16] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B 67 (2005) 301–-320.

[17] R. Genuer, J. Poggi, C. Tuleau-Malot, VSURF: an r package for variable selection using random forests, R J. 7 (2015) 19–33.

[18] J.L.M. Amaral, A.J. Lopes, J.M. Jansen, et al., Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease, Comput. Methods Programs Biomed. 105 (3) (2012) 183–193.

[19] E. Donald, Knuth: seminumerical algorithms, vol. 2 of The Art of Computer Programming, third edition, Addison-Wesley, 1998.

[20] P. Werga, G. Wallin, L. Skoog, B. Hamberger, Expanding role of fine-needle aspiration cytology in thyroid diagnosis and management, World J. Surg. 24 (2000) 907–912.

[21] C.G. Theoharis, K.M. Schofield, L. Hammers, R. Udelsman, D.C. Chhieng, The Bethesda thyroid fine-needle aspiration classification system: year 1 at an academic institution, Thyroid 19 (2009) 1215–1223.

[22] A. Mathur, J. Weng, W. Moses, S.M. Steinberg, R. Rahbari, M. Kitano, et al., A prospective study evaluating the accuracy of using combined clinical factors and candidate diagnostic markers to refine the accuracy of thyroid fine needle aspiration biopsy, Surgery 148 (1170-6) (2010) 1176–1177.

[23] S.T. Zahir, M. Vakili, A. Ghaneei, N.S. Sharahjin, F. Heidari, Ultrasound assistance in differentiating malignant thyroid nodules from benign ones, J. Ayub Med. Coll. Abbottabad 28 (2016) 644–649.

[24] M. Iqbal, Z. Mehmood, S. Rasul, H.S.S. Inamullah, I. Bokhari, Carcinoma thyroid in multi and uninodular goiter, J. Coll. Physicians Surg. 20 (2010) 310–312.

[25] M.C. Chammas, F.V. de Araujo, R.A. Moyses, M.D. Brescia, G.C. Mulatti, L.G. Brandao, et al., Predictive value for malignancy in the finding of microcalcifications on ultrasonography of thyroid nodules, Head Neck 30 (2008) 1206–1210.

[26] M. Rahimi, N. Farshchian, E. Rezaee, K. Shahebrahimi, H. Madani, To differentiate benign from malignant thyroid nodule comparison of sonography with FNAC findings, Pak. J. Med. Sci. 29 (2013) 77–80.

[27] E.K. Kim, C.S. Park, W.Y. Chung, K.K. Oh, D.I. Kim, J.T. Lee, et al., New sonographic criteria for recommending fine-needle aspiration biopsy of nonpalpable solid nodules of the thyroid, AJR Am. J. Roentgenol. 178 (2002) 687–691.

[28] Y.J. Choi, J.H. Baek, H.S. Park, et al., A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: initial clinical assessment, Thyroid 27 (4) (2017) 546–552.

[29] L.N. Li, J.H. Ouyang, H.L. Chen, D.Y. Liu, A computer aided diagnosis system for thyroid disease using extreme learning machine, J. Med. Syst. 36 (5) (2012) 3327–3337.

[30] S. Gitto, G. Grassi, C. De Angelis, et al., A computer-aided diagnosis system for the assessment and characterization of low-to-high suspicion thyroid nodules on ultrasound, Radiol. Med. (2018), https://doi.org/10.1007/s11547-018-0942-z [Epub ahead of print].

[31] E.Y. Jeong, H.L. Kim, E.J. Ha, et al., Computer-aided diagnosis system for thyroid nodules on ultrasonography: diagnostic performance and reproducibility based on the experience level of operators, Eur. Radiol. (2018), https://doi.org/10.1007/s00330-018-5772-9 [Epub ahead of print].

[32] 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, Med. Phys. 43 (1) (2016) 554.

[33] Y.E. Nikiforov, R.R. Seethala, G. Tallini, Z.W. Baloch, F. Basolo, L.D. Thompson, et al., Nomenclature revision for encapsulated follicular variant of papillary thyroid carcinoma: a paradigm shift to reduce overtreatment of indolent tumors, JAMA Oncol. 2 (2016) 1023–1029.