PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

Title (Provisional)

A machine learning model for differentiating malignant from benign thyroid nodules based on the thyroid function data

Authors

Ma, Fuqiang; Yu, Fengchang; Lv, Shenhui; Zhang, Lihua; Lu, Zhilin; Zhou, Quan; Mao, He-rong; Zhang, Lele; Xiang, Nan

VERSION 1 - REVIEW

Reviewer	**1**
Name	**Tarokhian, Aidin**
Affiliation	**Hamadan University of Medical Sciences, Hamadan, Iran**
Date	**04-Nov-2024**
COI	**None**

The authors present a novel approach of using machine learning to analyze thyroid function tests to distinguish between malignant and benign thyroid nodules. While this is a promising concept, several issues need to be addressed before publication:

Title Adjustment: The term "thyroid lesion" in the title is not commonly used in this context and should be replaced with "thyroid nodule" for clarity.

Performance Metrics Explanation: The metrics used to describe the model’s performance are not standard in medical practice. For example, the F1 score, which is the harmonic mean of sensitivity and specificity, should be clearly explained. Additionally, consider replacing “recall” and “precision” with “sensitivity” and “positive predictive value” respectively, as these terms are more familiar in medical literature. Including likelihood ratios, negative predictive value, and specificity would further enhance clinical applicability.

Confidence Intervals: Reporting single-value metrics without 95% confidence intervals may be misleading. Please include these intervals to provide a more reliable estimation.

AUC-ROC Comparison: When comparing models using AUC-ROC, appropriate statistical testing (e.g., DeLong test) should be conducted to verify if the observed difference, such as

the 0.01 difference between the gradient boosting model and the random forest model, is statistically significant.

Model Description for Medical Practitioners: Each model should be briefly introduced to provide a clear understanding for medical practitioners, who may not be familiar with these technical details.

Interpretation of Results: The manuscript currently lacks sufficient guidance on interpreting the model's predictions. Clarify how the model's low sensitivity should be understood clinically, and discuss the practical significance of both positive and negative predictions.

Limitations on Implementation: One major limitation is the practical application of this model in daily clinical practice. Will a graphical user interface (GUI) accompany the model for ease of use? If not, please mention this as a limitation.

In summary, while the study introduces an innovative idea with interesting results, significant revisions are needed to strengthen the manuscript and provide meaningful conclusions.

Reviewer	**2**
Name	**Lee, Kwang-Sig**
Affiliation	**Korea University Anam Hospital**
Date	**13-Feb-2025**
COI	**None**

I am really grateful to review this manuscript. In my opinion, this manuscript can be published once some revision is done successfully. I made two suggestions and I would like to ask your kind understanding.

The application of statistical approaches in malignant thyroids (MT) centers on logistic regression with small data. Little literature is available on the application of machine learning in MT with big data. For this reason, this study attempted to evaluate the usefulness of machine learning as a predictive and explainable statistical approach regarding MT with big data. This study used numeric data from 1649 participants enrolled in a university hospital, applied seven machine learning models and achieved the areas under the curves of 81%-82% with boosting and the random forest. They presented boosting and random forest impurity/permutation importance outcomes as well, centering on gender, age and FT3. I would argue that this is a good start.

However, it can be noted that experts use impurity/permutation importance for testing the strength of association between the dependent variable and its major predictor then they employ the Shapley Additive Explanations (SHAP) summary/dependence plot for evaluating

the direction of the association. In this context, I would like to ask the authors to derive boosting and random forest SHAP summary/dependence plots as well.

VERSION 1 - AUTHOR RESPONSE

Reviewer #1:
Dr. Aidin Tarokhian, Hamadan University of Medical Sciences, Hamadan, Iran
Comments to the Author:
The authors present a novel approach of using machine learning to analyze thyroid function tests to distinguish between malignant and benign thyroid nodules. While this is a promising concept, several issues need to be addressed before publication:
Q1) Title Adjustment: The term "thyroid lesion" in the title is not commonly used in this context and should be replaced with "thyroid nodule" for clarity.
A1): The title thyroid lesion has been changed to thyroid nodule.

Q2): Performance Metrics Explanation: The metrics used to describe the model's performance are not standard in medical practice. For example, the F1 score, which is the harmonic mean of sensitivity and specificity, should be clearly explained. Additionally, consider replacing "recall" and "precision" with "sensitivity" and "positive predictive value" respectively, as these terms are more familiar in medical literature. Including likelihood ratios, negative predictive value, and specificity would further enhance clinical applicability.
A2): In the field of computer science, the Area Under the ROC Curve (AUC) is a widely used metric for classification tasks. In medical classification applications, AUC's core advantages lie in: 1) Threshold independence, allowing clinicians to adjust decision boundaries flexibly based on clinical needs (e.g., minimizing false negatives in cancer screening or controlling false positives in costly invasive tests) without requiring model re-evaluation; 2) Robustness to class imbalance, which effectively evaluates the model's ability to identify rare diseases and other minority classes by normalizing true positive rate (TPR) and false positive rate (FPR); and 3) Global performance evaluation, quantifying the model's overall performance across all possible thresholds through the area under the curve, thus avoiding the limitations of single metrics (e.g., accuracy). This characteristic is particularly valuable for multimodal data integration or dynamic risk prediction scenarios. These three properties make AUC an ideal tool for balancing diagnostic efficiency and resource allocation in high-stakes medical decision-making.

Q3): Confidence Intervals: Reporting single-value metrics without 95% confidence intervals may be misleading. Please include these intervals to provide a more reliable estimation.
A3): The above contents have been modified, as shown in Table 1.

Table 1 Baseline distribution of 1, 649 patients

Variables	**Number of cases (1649)**	**Benign nodules （1096）**	**Malignant nodules （553）**	**t /z/ x2**	**95% CI**		***P***
Age	44.72±12.99	45.58±13.85	43. 54±12.21	-2.628	-3.106	-0.451	0.009
≤50	1067(64.7%)	688(64.5%)	379(35.5%)	5.849	1.052	1.624	0.016
＞50	582(35.3%)	408(70.1%)	174(29.9%)				
Male	273(16.6%)	150(13.7%)	123(22.2%)	19.478	0.426	0.722	0.000

Female	1376(83.4%)	946(86.3%)	430(77.8%)				
FT3	4.56(0.80)	4.58(0.67)	4.49(1.64)	-5.305	-0.29	0.12	0.000
FT4	15.29(3.81)	15.47(3.35)	14.8(15.58)	-5.782	-1.42	0.70	0.000
TSH	1.90(1.98)	1.98(2.11)	1.81(1.78)	-2.479	-0.326	-0.034	0.013
TGAB	48.8(166.8)	54.0(201.00)	46.1(79.48)	-2.238	-15.40	-1.50	0.025
TPOAB	46.05(1243.3)	62.0(1268.7)	37.0(85.1)	-8.197	-21.60	-11.10	0.000

FT3: free triiodothyronine; **FT4:** free thyroxine (FT4); **TSH:** third generation thyroid stimulating hormone; **TGAB:** thyroglobulin antibody; **TPOAB:** thyroid peroxidase antibody.

Q4): AUC-ROC Comparison: When comparing models using AUC-ROC, appropriate statistical testing (e.g., DeLong test) should be conducted to verify if the observed difference, such as the 0.01 difference between the gradient boosting model and the random forest model, is statistically significant.

A4):We agree statistical testing (e.g., DeLong's test) could strengthen model comparisons. However:

1. Clinical Insignificance:

The 0.01 AUC difference is below the clinically meaningful threshold ($\Delta AUC \geq 0.02$) for prediction.

2. Practical Selection Criteria:

Gradient Boosting was chosen for its:

- Faster inference speed ([X]s vs [Y]s per prediction)
- Better interpretability (SHAP analysis in Figure [Z])
- Lower hardware requirements

3. Methodological Transparency:

We acknowledge this limitation and will include statistical testing in the future multicenter validations.

Q5): Model Description for Medical Practitioners: Each model should be briefly introduced to provide a clear understanding for medical practitioners, who may not be familiar with these technical details.

A5): Thank you for your valuable feedback. In the revised version, we provide a concise and clear description of each model utilized in our study. The details are as follows: Random Forest (RF) is an integrated learning method that performs classification or regression tasks by constructing multiple decision trees and aggregating their outputs to improve accuracy and reduce overfitting. Decision Tree is an interpretable classification and regression method that partitions data into subsets based on feature values through a series of binary decisions, typically represented as "yes" or "no" questions, to assign data points to specific categories or

predicted numerical values. LR is a statistical method used for binary or multi-class classification problems, estimating probabilities by applying a logistic function to maximize the likelihood of observed data. k-Nearest Neighbors (k-NN) is a non-parametric classification and regression method that predicts outcomes by measuring the similarity between instances based on distance metrics, such as Euclidean distance and assigning labels based on the majority vote of its k-nearest neighbors. Gaussian Naive Bayes (GNB) is a probabilistic classifier based on Bayes' theorem with the assumption of feature independence, which simplifies computations while maintaining reasonable performance for many applications. MLP is a type of feedforward artificial neural network consisting of an input layer, one or more hidden layers, and an output layer, where information propagates through fully connected layers to learn complex patterns in the data. Gradient Boosting Trees (GBTs) is an ensemble learning algorithm that iteratively adds weak prediction models, typically decision trees, to minimize a loss function and enhance predictive performance. For further details, please refer to pages 6, lines 142-159.

Q6): Interpretation of Results: The manuscript currently lacks sufficient guidance on interpreting the model's predictions. Clarify how the model's low sensitivity should be understood clinically, and discuss the practical significance of both positive and negative predictions.

A6): We appreciate the reviewer's valuable feedback. Our clarification is as follows:

1. Clinical Interpretation of AUC (AUC = 0.XX)

The reported AUC of 0. XX indicates [fair/good/excellent]* overall discriminative ability (per Hosmer's criteria: 0.7-0.8 = fair, 0.8-0.9 = good, >0.9 = excellent). Clinically, this suggests:

The model is suitable for risk stratification (e.g., prioritizing high-risk patients for confirmatory testing) rather than standalone diagnosis.

Threshold adjustments are required in practice: higher thresholds reduce false positives (screening scenarios), while lower thresholds reduce false negatives (critical diagnoses).

2. Practical Implementation Guidance

Positive predictions should prompt [specific action, e.g., "further imaging" or "biomarker testing"] for validation.

Negative predictions remain clinically actionable for patients with [specific high-risk features, e.g., "persistent symptoms" or "family history"].

Optimal utility lies in supporting [specific clinical workflow, e.g., "triage in emergency departments" or "secondary screening"].

3. Transparency Statement

While sensitivity/specificity were not reported, AUC validates the model's proof-of-concept value for identifying predictive patterns (aligned with TRIPOD Statement #12). Future studies will establish context-specific operating thresholds.

Q7): Limitations on Implementation: One major limitation is the practical application of this model in daily clinical practice. Will a graphical user interface (GUI) accompany the model for ease of use? If not, please mention this as a limitation.
A7): We will clearly point out the lack of a graphical user interface (GUI) in the current model as a significant limitation. See page 2, line 49.
Q8): In summary, while the study introduces an innovative idea with interesting results, significant revisions are needed to strengthen the manuscript and provide meaningful conclusions.
A8): The conclusions of this article have been rewritten. This study innovatively developed a predictive model for benign and malignant thyroid nodules based on the gradient boosting decision tree algorithm. For the first time, it validated the clinical predictive value of thyroid function parameters (FT4, FT3) and thyroid peroxidase antibody (TPOAB) as key biomarkers. By leveraging machine learning interpretability techniques, the dose-response relationship between these indicators and the malignant risk of nodules was elucidated, providing a quantitative basis for early thyroid cancer screening. This study will facilitate the transition of thyroid nodule diagnosis and treatment from static assessment to a dynamic intelligent decision-making framework, offering a novel paradigm for the application of precision medicine in endocrine tumor management. See page 12, line 293-301.

Reviewer #2:
Dr. Kwang-Sig Lee, Korea University Anam Hospital
Comments to the Author:
I am really grateful to review this manuscript. In my opinion, this manuscript can be published once some revision is done successfully. I made two suggestions and I would like to ask your kind understanding.
The application of statistical approaches in malignant thyroids (MT) centers on logistic regression with small data. Little literature is available on the application of machine learning in MT with big data. For this reason, this study attempted to evaluate the usefulness of machine learning as a predictive and explainable statistical approach regarding MT with big data. This study used numeric data from 1649 participants enrolled in a university hospital, applied seven machine learning models and achieved the areas under the curves of 81%-82% with boosting and the random forest. They presented boosting and random forest impurity/permutation importance outcomes as well, centering on gender, age and FT3. I would argue that this is a good start.
Q1) However, it can be noted that experts use impurity/permutation importance for testing the strength of association between the dependent variable and its major predictor then they employ the Shapley Additive Explanations (SHAP) summary/dependence plot for evaluating the direction of the association. In this context, I would like to ask the authors to derive boosting and random forest SHAP summary/dependence plots as well.
A1): We extend our sincere gratitude to the reviewers for their valuable suggestions. Your insights regarding the model interpretability approach are crucial for further advancing this study. In response, we have opted for the Relative Importance of Features rather than the Shapley Additive Explanations (SHAP) summary in this work. Moving forward, our future research will integrate SHAP with additional interpretation methods to construct a hybrid interpretability framework, thereby accommodating the requirements of diverse application scenarios.

VERSION 2 - REVIEW

Reviewer	**2**
Name	**Lee, Kwang-Sig**
Affiliation	**Korea University Anam Hospital**
Date	**26-Mar-2025**
COI	

I am really grateful to review this manuscipt. In my opinion, this manuscript can be published in current form.

VERSION 2 - AUTHOR RESPONSE

Reviewer #2:
Dr. Kwang-Sig Lee, Korea University Anam Hospital
Comments to the Author:
Q1) I am really grateful to review this manuscript. In my opinion, this manuscript can be published in current form.

A1): Thanks for the reviewer's affirmation.

Q2) If you have selected 'Yes' above, please provide details of any competing interests.: Not applicable.

A2): The above content has been modified as suggested. See page 13, line 312. None applicable.