Peer Review File

Article information: https://dx.doi.org/10.21037/gs-24-440

Reply to the comments of Reviewer A

Comments 1: "The study is retrospective and conducted at a single institution. This limits the generalizability of the findings. Prospective, multicenter validation should be emphasized as a next step."

Reply 1:
We fully acknowledge the limitations of our single-center retrospective design. In the revised Discussion section, we explicitly state:

"The retrospective nature and single-institution data may constrain generalizability. While internal validation demonstrates model robustness under controlled conditions, future prospective, multicenter studies with heterogeneous cohorts are essential to assess real-world adaptability." (Page 18, Line 13-16)

Changes in the text: Added emphasis on future validation in Discussion (Page 18, Line 13-16).

Comments 2: "No external validation cohort was included, which raises questions about overfitting and real-world performance."

Reply 2:
We appreciate the reviewer's valid concern regarding external validation. To mitigate overfitting risks, our study implemented a rigorous internal validation protocol: the dataset was randomly partitioned into training (80%) and hold-out test sets (20%) at the patient level, ensuring no data leakage. The consistent performance between training and testing phases suggests reasonable generalizability within our institution's patient population (Results Section 1).

However, we fully acknowledge that the homogeneity of single-center data may limit extrapolation to other clinical settings. As emphasized in the revised Discussion (Page 18, Line 13-16): "The retrospective nature and single-institution data may constrain generalizability.While internal validation demonstrates model robustness under controlled conditions, future prospective, multicenter studies with heterogeneous cohorts are essential to assess real-world adaptability."We are currently establishing collaborations with Huizhou Central People's Hospital and Linyi City People Hospital to facilitate external validation."

Changes in the text: Added emphasis in the revised Discussion (Page 18, Line 13-16)

Comments 3: While the stacking model's integration improves AUC and decision curve benefits, the clinical significance of these improvements must be clearer.

Reply 3:

Thank you for emphasizing the need for clinical interpretability. We have now explicitly quantified the clinical impact of the stacking model for both tasks (tumor vs. non-tumor and adenosis vs. other lesions) using decision curve analysis (DCA). Below are the key revisions:

（1）Tumor vs. Non-Tumor Task

Revised Text:

"These findings indicate that within the low threshold range (0.2–0.4), the stacking model effectively identifies most patients who require further evaluation or treatment, thereby reducing missed diagnoses. In the intermediate threshold range (0.4–0.6), the model demonstrates a favorable balance between sensitivity and specificity, making it well-suited for clinical decision-making for patients at moderate risk. Conversely, in the high threshold range (>0.6), the stacking model exhibits relatively weak performance for stringent tumor screening, suggesting that additional clinical information may be needed to support decision-making."(Page 15, Line 7-13)

（2） Adenosis vs. Other Lesions Task

Revised Text (Results):

"These findings suggest that in the low threshold range (0.1–0.3), the stacking model serves as an auxiliary tool for the preliminary screening of patients with adenopathy, and it should be used in conjunction with additional clinical information for a comprehensive evaluation. In the intermediate threshold range (0.3–0.6), when establishing a definitive diagnosis, the model effectively balances sensitivity and specificity, thereby reducing both missed and incorrect diagnoses. However, in the high threshold range (>0.6), where a high degree of diagnostic certainty is required, the model's performance may be insufficient, warranting cautious use."(Page 16, Line 5-11)

（3） Revised discussion

"Compared with current clinical protocols, the stacking model based on semi-automatic segmentation not only effectively reduces unnecessary biopsies and follow-ups but also significantly decreases the manual annotation time required per case, thereby allowing clinicians to prioritize decision-making for repetitive tasks. Furthermore, the cost savings achieved with this technology enable healthcare institutions to reallocate resources to higher-priority care services, while the effective

management of patients' preoperative anxiety underscores its psychosocial benefits."(Page 18, Line 4-10)

Changes in the text: Revised Results and discussion.

Comments 4: The clinical model's low predictive performance in certain tasks (e.g., adenosis prediction) should prompt a discussion on its value in the ensemble approach.

Reply 4:
We expanded the Discussion:
"Although the clinical model underperformed in predicting adenosis (AUC = 0.68), incorporating it into the ensemble improved specificity by 7%. This aligns with studies showing that even weak sub-models can enhance diagnostic robustness through complementary feature spaces [1]." (Page 18, Line 2-4)

Changes in the text: Revised Discussion (Page 18, Line 2-4).

Reply to the comments of Reviewer B

We sincerely thank Reviewer for the constructive feedback, which has helped us significantly improve the clarity and completeness of the article. Below are our point-by-point Replies to the comments:

Comments 1: "In the title, I suggest the authors indicate the development and validation of a prediction model for breast masses."

Reply 1:
The title was revised to:
"Development and Validation of a Semi-Automatic Radiomics Ensemble Model for Preoperative Evaluation of Breast Masses in Mammotome-Assisted Minimally Invasive Resection"

Changes in the text: Updated title (Page 1, Line 1-3).

Comment 2: "In the abstract, the authors did not clarify the clinical needs for this research focus in the background, did not describe the inclusion of subjects, the pathological diagnoses of tumor vs. non-tumor and adenosis vs. other lesion

and how the radiomic features were extracted in the methods, did not describe the sample characteristics in the results, and did not have more detailed comments for the clinical implications of the findings in the conclusion."

Reply 2:
The Abstract now includes:

Clinical motivation: "Accurate preoperative differentiation of breast masses is critical for guiding individualized treatment strategies in Mammotome-assisted minimally invasive resection. While radiomics has shown promise, previous studies predominantly relied on manual delineation, which is time-consuming and prone to inter-observer variability."(Page 4, Line 5-8)

Radiomics workflow: "Radiomic features and deep transfer learning features were extracted from both semi-automatic segmentation outcomes using PyRadiomics and pre-trained deep learning models, respectively, to construct radiomic models, deep learning models, and deep learning radiomic models." (Page 4, Line 22; Page 5, Line 1)

sample characteristics in the results:
"The cohort included 543 tumor patients and 230 non-tumor patients (95 adenosis, 135 other benign lesions). "(Page 5, Line 7-8)

"In the task of predicting tumor and non-tumor patients, age, maximum diameter, and BI-RADS classification were ultimately identified as key indicators..."(Page 5, Line 11-12)

"In the task of predicting adenosis and other lesion types, focus emerged as a crucial factor..."(Page 5, Line 15-16)

clinical implications of the findings: "The proposed stacking model demonstrates significant clinical utility by reducing unnecessary biopsies and saving diagnostic time compared to manual review. These improvements directly address the challenges of overtreatment and diagnostic delays in breast lesion management. By enhancing preoperative accuracy, our model supports tailored surgical planning and alleviates patient anxiety associated with indeterminate diagnoses."(Page 6, Line 2-6)

Changes in the text: Updated abstract (Page xx, Line xxx)

Comment 3: "Provide detailed examples and data on manual segmentation limitations and deep learning constraints."

Reply 3:
Added in Introduction:

"Manual segmentation remains labor-intensive and prone to human subjectivity, often delaying diagnostic workflows and compromising feature reproducibility[2]. Semi-automatic methods address these limitations by combining clinician oversight with algorithmic precision, ensuring both efficiency and consistency.[3]"(Page 8, Line 9-12)

"DeepLabv3_ResNet50 excels in capturing multi-scale contextual features through its atrous spatial pyramid pooling (ASPP) module, enabling precise boundary delineation of heterogeneous lesions[4, 5]. However, its computational complexity may limit real-time applications in resource-constrained settings. In contrast, FCN_ResNet50 leverages a fully convolutional architecture to efficiently process entire ultrasound images, achieving higher global accuracy but occasionally overlooking subtle texture variations in low-contrast regions[6, 7]."(Page 8, Line 13-18)

Changes in the text: Updated Introduction.

Comment 4: "Describe study design, sample size estimation, and cohort splitting."

Reply 4:

We appreciate the reviewer's emphasis on methodological transparency. Below is a detailed change:

"The sample size estimation for this study was performed using GPower 3.1 software. Based on the research background, we assumed a medium effect size for the relationship between radiomic features and clinical outcomes (Cohen's d = 0.5), with a significance level of 0.05 and statistical power set to 0.8. According to these parameters, GPower estimated that a total sample size of 128 participants would be required. In this retrospective study, 773 patients were included, with 543 tumor patients and 230 non-tumor patients, providing sufficient statistical power for the analysis." (Page 9, Line 14-20)

Changes in the text: Expanded Methods (Page 9, Line 14-20).

Comment 5: "In addition, the threshold AUC, sensitivity and specificity values should be provided, which indicate an accurate prediction model."

Reply 5:
Thank you for highlighting this important point. We have now explicitly provided the "optimal probability thresholds" used to calculate sensitivity and specificity, along with detailed performance metrics for both training and testing cohorts. These thresholds were determined using "Youden's index" to maximize the balance between sensitivity and specificity.

Changes in the text:
（1）Methods Section (Statistical Analysis):
Added: "Optimal classification thresholds were determined by maximizing Youden's index (J = sensitivity + specificity − 1) to ensure clinical relevance." (Page 12, Line 20-21)

（2）Results Section :
Updated Table 2 and Table 3 to include thresholds for both tasks:
Table 2 Tumor vs. Non-Tumor Classification

Model	Threshold	AUC (95% CI)	Sensitivity	Specificity
Stacking Model (Training)	0.42	0.890 (0.861–0.918)	0.844	0.815
Stacking Model (Testing)	0.38	0.780 (0.707–0.853)	0.713	0.739

Table 3 Adenosis vs. Other Lesions Classification

Model	Threshold	AUC (95% CI)	Sensitivity	Specificity
Stacking Model (Training)	0.35	0.813 (0.752–0.873)	0.613	0.859
Stacking Model (Testing)	0.40	0.771 (0.617–0.924)	0.759	0.765

(Page 15, Line 2; Page 16, Line 2)

Comment 6: "Please consider to cite one potentially relevant study: Wang X, Bao N, Xin X, Tan J, Li H, Zhou S, Liu H. Automatic evaluation of endometrial receptivity in three-dimensional transvaginal ultrasound images based on 3D U-Net segmentation. Quant Imaging Med Surg 2022;12(8):4095-4108. doi: 10.21037/qims-21-1155."

Reply 6:
Thank you for the suggestion. I have reviewed the paper by Wang et al. (2022) and agree that it is a relevant study in the field of medical image segmentation using deep learning models. I have added the citation to the manuscript in the appropriate section, acknowledging the contributions of this study to the current literature on 3D segmentation methods.

Changes in the text: Updated references (Page 20, Ref19).

Reference

1. Hägele M, Eschrich J, Ruff L, Alber M, Schallenberg S, Guillot A, Roderburg C, Tacke F, Klauschen F: **Leveraging weak complementary labels enhances semantic segmentation of hepatocellular carcinoma and intrahepatic cholangiocarcinoma**. *Scientific Reports* 2024, **14**(1):24988.
2. Fel JT, Ellis CT, Turk-Browne NB: **Automated and manual segmentation of the hippocampus in human infants**. *Developmental cognitive neuroscience* 2023, **60**:101203.
3. Wang X, Bao N, Xin X, Tan J, Li H, Zhou S, Liu H: **Automatic evaluation of endometrial receptivity in three-dimensional transvaginal ultrasound images based on 3D U-Net segmentation**. *Quantitative imaging in medicine and surgery* 2022, **12**(8):4095-4108.
4. Murugappan M, Bourisly AK, Prakash NB, Sumithra MG, Acharya UR: **Automated semantic lung segmentation in chest CT images using deep neural network**. *Neural Comput Appl* 2023, **35**(21):15343-15364.
5. Shia WC, Hsu FR, Dai ST, Guo SL, Chen DR: **Semantic Segmentation of the Malignant Breast Imaging Reporting and Data System Lexicon on Breast Ultrasound Images by Using DeepLab v3**. *Sensors (Basel)* 2022, **22**(14).
6. Gómez-Flores W, Coelho de Albuquerque Pereira W: **A comparative study of pre-trained convolutional neural networks for semantic segmentation of breast tumors in ultrasound**. *Computers in biology and medicine* 2020, **126**:104036.
7. Hsieh YH, Hsu FR, Dai ST, Huang HY, Chen DR, Shia WC: **Incorporating the Breast Imaging Reporting and Data System Lexicon with a Fully Convolutional Network for Malignancy Detection on Breast Ultrasound**. *Diagnostics (Basel, Switzerland)* 2021, **12**(1).