



Dear Reviewer,

Thank you for taking the time to review our manuscript (Manuscript ID: 106103) and providing valuable feedback. We appreciate your careful consideration of our work and your thoughtful comments. We have made improvements to the revised document based on your feedback and request. In addition, we would like to present the following information:

Reviewer:

The manuscript presents a retrospective study employing CT-based deep learning (DL) radiomics for the non-invasive preoperative evaluation of the tumor immune microenvironment (TIME) in colorectal cancer (CRC). The proposed DL models demonstrate promising predictive performance in assessing key immune-related biomarkers, such as the tumor-stroma ratio (TSR), tumor-infiltrating lymphocytes (TILs), and immune score (IS). While the study is scientifically significant, certain aspects require further refinement to improve clarity, reproducibility, and clinical applicability. For Methods:

1. Feature Importance & Model Interpretation The manuscript lacks a detailed analysis of the most influential features contributing to model performance. It would be valuable to include feature importance analysis using SHAP (Shapley Additive Explanations) or other interpretable AI techniques to better understand which radiomics or DL-derived features drive model predictions.



2. Justification for DenseNet-169 Selection While the manuscript compares multiple deep learning architectures, it does not provide a clear justification for why DenseNet-169 was ultimately selected. It appears that AUC values were used for comparison, but the reasoning should be made more explicit. For instance, were other performance metrics (e.g., sensitivity, specificity, calibration errors) also considered?

3. Confidence Interval Overlaps Although DenseNet-169 achieved the highest AUC, its confidence interval (CI) overlaps with those of other models. This raises questions about whether the difference is statistically significant. The authors should explicitly discuss this overlap and clarify whether the performance superiority is robust or potentially due to sample variation.

4. Segmentation Consistency & Interobserver Variability The CT segmentation and labeling process is well described, but interobserver variability among radiologists is not sufficiently addressed. Reporting inter-rater agreement metrics such as Dice similarity coefficient or Cohen's kappa would help demonstrate annotation consistency and validate the reliability of tumor region delineation.

5.

5.1 Training vs. Validation Performance Discrepancies In Table 2, some models exhibit better performance in the validation set than in the training set, which is unusual because training typically involves replicating known data and



should, in theory, achieve a "perfect" AUC. This discrepancy raises concerns about potential issues in hyperparameter tuning, overfitting, or data leakage. Possible reasons for this observation should be discussed. Moreover, if hyperparameter optimization was conducted, the methodology should be clearly described.

For Tables & Figures:

- Table 1 - Statistical Methods Clarification** The table legend should explicitly describe the statistical methods used for categorical and numeric variables to enhance clarity. Additionally, the reference to "a" is missing and should be checked.
- Formatting in Tables & Text** There should be a space between numbers and parentheses throughout the manuscript (e.g., change "2(1.5,3.2)" to "2 (1.5, 3.2)"). This issue is present in multiple sections and should be systematically corrected.
- Figure 1 - Clarity & Labeling** - The resolution (DPI) of Figure 1 should be improved for better readability. - For the deep learning procedure visualization, the X-axis and Y-axis must be clearly labeled to indicate what is being measured. - The authors should explicitly state that subsequent plots (e.g., Grad-CAM visualizations, ROC curves) correspond to the best-performing model. However, it would be preferable to first showcase model development and comparisons, followed by the performance of the selected model.

For Writing & Formatting & Grammar & Typographical Errors:

- There are several minor typos throughout the manuscript. For example, "CRC is one the main causes..." → should be "CRC



is one of the main causes...". 2. In citation formatting, ensure proper spacing after periods (e.g., "[1].It is anticipated..." should be "[1]. It is anticipated...").

The author's answer:

1.Dear Reviewer,We sincerely appreciate your suggestion. Indeed, both radiomics and deep learning have numerous features. However, deep learning features are based on models and training tasks, and do not carry real-world significance. Our model is based on a deep learning approach trained on images. The model's visualization is presented using Class Activation Mapping (CAM), which works by weighted combinations of feature maps from convolutional layers to locate key regions associated with specific categories. The details of this are shown in Figure 5 of the manuscript. Therefore, we have only analyzed the feature importance of radiomics. For example, in a CT-based radiomics model predicting immune scores, the feature importance is illustrated in the figure below. Since the primary focus of this study is on deep learning-based training results, we did not present the feature importance analysis based on traditional machine learning features in the manuscript.

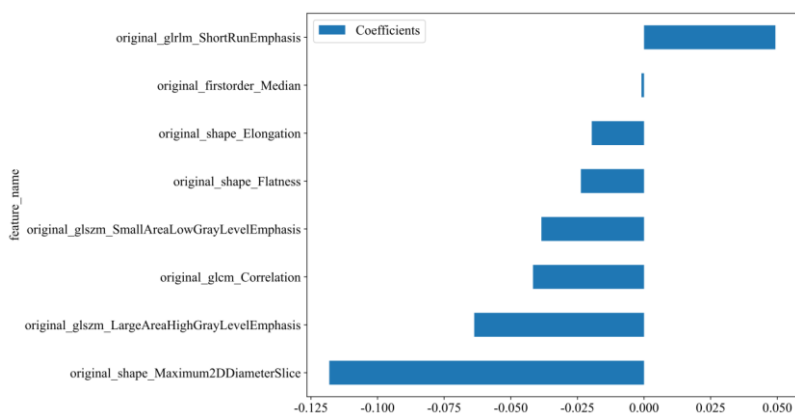


Figure: Analysis of the importance of radiomics features.



2.ROC curve analysis, as a widely used evaluation method for binary classification models with excellent visualization features, has been extensively applied. In this study, based on the performance of the test set of the developed model and the AUC values of the ROC curves for both the training and test sets, we selected DenseNet169, which demonstrated the best predictive performance, for display. Additionally, in the table below, we also present the specific parameters of other models (e.g., sensitivity, specificity, PPV, NPV). We observed that the AUC value of DenseNet169 is the closest.

Table: Detailed information on each deep learning model

ModelName	ACC	AUC	95%CI	Sensitivity	Specificity	PPV	NPV	Precision	Recall	F1	Threshold	Cohort
0 densenet121	0.741	0.797	0.7383-0.8556	0.707	0.769	0.714	0.762	0.714	0.707	0.711	0.520	Train
1 densenet121	0.800	0.759	0.6502-0.8676	0.705	0.882	0.838	0.776	0.838	0.705	0.765	0.407	Test
2 densenet169	0.709	0.780	0.7185-0.8406	0.768	0.661	0.650	0.777	0.650	0.768	0.704	0.385	Train
3 densenet169	0.768	0.772	0.6741-0.8696	0.682	0.834	0.789	0.754	0.789	0.682	0.732	0.541	Test
4 densenet201	0.718	0.765	0.7018-0.8274	0.495	0.901	0.803	0.686	0.803	0.495	0.612	0.591	Train
5 densenet201	0.768	0.737	0.6305-0.8432	0.636	0.882	0.824	0.738	0.824	0.636	0.718	0.586	Test
6 resnet101	0.786	0.852	0.8032-0.9011	0.859	0.727	0.720	0.863	0.720	0.859	0.783	0.432	Train
7 resnet101	0.737	0.752	0.6503-0.8541	0.636	0.824	0.757	0.724	0.757	0.636	0.691	0.446	Test
8 resnet152	0.732	0.816	0.7603-0.8718	0.869	0.620	0.652	0.852	0.652	0.869	0.745	0.336	Train
9 resnet152	0.737	0.736	0.6272-0.8438	0.614	0.843	0.771	0.717	0.771	0.614	0.684	0.501	Test
10 resnet34	0.782	0.860	0.8120-0.9072	0.808	0.760	0.734	0.829	0.734	0.808	0.769	0.444	Train
11 resnet34	0.747	0.741	0.6344-0.8474	0.705	0.784	0.738	0.755	0.738	0.705	0.721	0.498	Test
12 resnet50	0.795	0.863	0.8144-0.9120	0.828	0.769	0.745	0.845	0.745	0.828	0.785	0.464	Train
13 resnet50	0.747	0.754	0.6491-0.8598	0.614	0.863	0.794	0.721	0.794	0.614	0.692	0.501	Test

3.Dear Reviewer, We greatly appreciate your suggestion. Therefore, using the deep learning model for predicting immune scores as an example, we conducted a Dolong test on different deep models, as shown in the table below ($P \leq 0.05$, statistically significant). The results indicate that there is no significant statistical difference between DenseNet-169 and DenseNet-121, DenseNet-201, ResNet34, ResNet50, and ResNet101. As a result, we selected the DenseNet-169 model, which performed better



in terms of AUC on the test set and had a closer AUC between the training and test sets, as the model for display.

Table 2 : Paired sample differences in AUC

Model Name	z	P value ^a	AUC difference	SE difference ^b	95% CI	
					lower bound	Upper Bound
densenet169-densenet121	0.295	0.768	0.008	0.229	-0.046	0.063
densenet169-densenet201	0.672	0.502	0.019	0.230	-0.037	0.075
densenet169-resnet34	-1.804	0.071	-0.046	0.222	-0.095	0.004
densenet169-resnet50	-1.638	0.101	-0.046	0.224	-0.101	0.009
densenet169-resnet101	-1.715	0.086	-0.046	0.221	-0.099	0.007

a. Null hypothesis: True regional difference = 0; b. Based on non-parametric assumptions, respectively; SE: standard error; CI: confidence interval.

4. Dear Reviewer, We greatly appreciate your suggestion. In order to ensure the reproducibility of the radiomics and deep learning models, it became essential to assess the consistency of the delineations made by the two annotating physicians. Therefore, we randomly selected 11 features extracted from CT images delineated by the two physicians and performed a Mann-Whitney U test on the two sets of features. The results showed that there were no statistically significant differences between the two sets of features ($p < 0.05$) (as shown in the table below). This, to some extent, demonstrates that the regions of interest (ROIs) annotated by the two radiologists exhibit consistency (Page 11 of the text marked in yellow).

Table: Differential analysis of ROI extraction features based on two physician markers

Features	Radiologist	Radiologist	Z Value	P Value
	1(n=30)	2(n=30)		
original_firstorder_10Percentile	-0.11(-0.61,0.65)	-0.10(-0.60,0.65)	-0.030	0.976
original_firstorder_Kurtosis	-0.45(-0.65,0.85)	-0.44(-0.66,0.88)	-0.030	0.976
original_firstorder_Median	-0.21(-0.77,0.90)	-0.21(-0.77,0.90)	-0.030	0.976
original_firstorder_Skewness	-0.21(-0.68,0.88)	-0.21(-0.68,0.88)	-0.044	0.965
original_firstorder_TotalEnergy	-0.29(-0.60,0.10)	-0.29(-0.60,0.11)	-0.044	0.965
original_glcm_ClusterShade	-0.08(-0.54,0.77)	-0.08(-0.54,0.77)	-0.015	0.988
original_glcm_Idn	0.18(-0.48,0.63)	0.18(-0.47,0.63)	-0.015	0.988



original_glcm_Imc2	0.27(-0.07,0.64)	0.27(-0.06,0.65)	-0.030	0.953
original_glcm_MCC	-0.14(-0.69,0.89)	-0.07(-0.77,0.98)	-0.030	0.953
original_glcm_MaximumProbability	-0.18(-0.75,0.69)	-0.17(-0.78,0.63)	-0.030	0.976
original_glrlm_GrayLevelNonUniformity	-0.31(-0.46,-0.06)	-0.25 (-0.71,0.57)	-0.192	0.848

5. Dear Reviewer, We have discussed the relevant issue regarding some models showing better performance in the validation set compared to the training set, which may be due to the dataset division. To address this, we could re-allocate the dataset or perform 10-fold cross-validation in the future to increase the scientific and objective rigor of the model (**highlighted sections on pages 16-17 of the manuscript**). Additionally, Figure 1 and the corresponding Figure 4 have been revised to include the resolution and coordinate labels for the deep learning training process' loss curve (showing the fluctuation of loss as the number of iterations increases; the x-axis represents iterations, while the y-axis does not indicate the specific loss values. For display purposes, we have applied translation and stretching to different curves).

Regarding the development and comparison of the displayed model, considering the length of the manuscript and the focus on the key research, we only present the performance of DenseNet121 and DenseNet169. We are happy to provide details on the development of other models upon request. Additionally, we have revised the format of the figures and tables, as well as corrected some writing errors based on your feedback, in order to meet the publication requirements.

Once again, thank you for your constructive feedback, which has helped us to improve the quality of our manuscript. We have carefully considered all of your suggestions and made necessary revisions to the manuscript.



We appreciate your time and effort in reviewing our work. Please let us know if you have any further comments or concerns.

Sincerely,

Prof. Dr. Mingxu Da