Accepted manuscript available at: https://doi.org/10.1109/MetroXRAINE62247.2024.10797090

# Low-Dose CT-Based Radiomic Model to Identify Malignant Lung Nodules

Jiaying Liu, Anna Corti, Valentina D.A. Corino, Luca Mainardi
Department of Electronics, Information and Bioengineering
Politecnico di milano, Milan, Italy
{jiaying.liu, anna.corti, valentina.corino, luca.mainardi}@polimi.it

***Abstract*— Lung cancer screening using low-dose computed tomography (LDCT) faces challenges in nodule classification and radiologist burden. This study introduces a novel approach leveraging degraded standard-dose CT (SDCT) images and develops an explainable radiomics-based model to improve lung nodule classification. Using data from LIDC-IDRI public dataset, our pipeline extracts and preprocesses 851 radiomic features per nodule, ensuring stability, discriminant characteristics, and non-redundancy. To address dataset imbalance, we applied data augmentation and random undersampling techniques. Varying the number of features selected by gradient-boosted decision tree across multiple classifiers, we identified an optimal combination of 15 features with random forest classifier, achieving balanced accuracy, sensitivity, specificity and AUC-ROC scores of 0.82 ± 0.04, 0.81 ± 0.08, 0.83 ± 0.05, 0.88 ± 0.05, respectively. Explainability analysis using SHAP highlights features influencing classification, including *major axis length*, *sphericity*, and *wavelet_LHH_ngtdm_Busyness*. In conclusion, our study demonstrates the potential of integrating degraded SDCT and radiomics to improve lung nodule classification.**

***Keywords*— *radiomics, low-dose computed tomography (LDCT), lung cancer, nodule classification.***

## I. INTRODUCTION

Lung cancer is the deadliest malignancy in the world, being responsible for 18% of total cancer deaths in 2020 [1]. While low-dose computed tomography (LDCT) screening has shown promise in reducing mortality rates compared to X-rays [2], the prevalence of false positive rates, reaching 96.4% of the positive screening, poses a significant challenge [2]. As screening programs are being implemented in the US and ongoing trials in Europe are assessing the benefits and cost-effectiveness of lung cancer screening, the workload on radiologists increases significantly. To overcome these challenges, extensive research has focused on developing objective and efficient diagnostic tools, with radiomics and deep learning emerging as promising methods [3].

Radiomics is based on the premise that image-captured characteristics of tumors, such as shape and texture, are potential tumor biomarkers that can be used to stratify patients, predict patient clinical outcomes or further characterize tumor phenotypes.

The development of radiomics-based models requires annotated datasets, which are currently lacking in the context of lung screening LDCT images. For instance, the extensive National Lung Screening Trial (NLST) contains LDCTs from 53,000 participants but lacks segmented nodules. On the other hand, the Lung Imaging Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [4] consists of diagnostic and lung cancer screening CT scans with marked-up annotated lesions by experienced radiologists. However, the number of LDCT images is relatively limited (315 out of the 1010 CTs, considering images acquired with tube current < 80 mAs). To address this challenge, we leverage the annotations from the LIDC-IDRI dataset by degrading standard-dose CT (SDCT) scans. This degradation step was essential to ensure compatibility between the different image types and facilitate the development of a classifier capable of effectively analyzing both SDCT and LDCT modalities.

Focusing on radiomics with machine learning studies applied to pulmonary nodule prediction, many have been shown to be promising in previous studies[5]. However, some considered small datasets ([6], [7]) and others achieved low sensitivity ([8], [9], [10]). For instance, Choi et al. [6] used two radiomics features and trained a support vector machine, obtaining scores of 0.89 and 0.87 for AUC and sensitivity. However, [6] only included 72 nodules in their study. On the other hand, Alahmari et al. [9] considered 467 patients in their study but obtained a low sensitivity of 0.56. Causey et al. [8] trained a logistic regression model using only one feature (square root of largest cross-sectional area) and obtained AUC score of 0.94, but also a relatively low sensitivity of 0.70. Likewise, Rundo et al. [10] obtained a sensitivity of 0.67 considering 32 radiomic features and logistic regression model. Other studies deliberately selected balanced datasets to address class imbalance [8], [11]. None of the aforementioned studies used degraded SDCT images to simulate LDCT.

The aim of this study is to build a novel explainable radiomic model to distinguish between likely benign and likely malign lung nodules, as annotated by radiological assessment, addressing the limitations associated with different dosage CT images, imbalanced dataset and false positive rates.

## II. METHODS

The pipeline (Fig. 1) consists of data partition, image preprocessing, feature extraction and preprocessing, feature selection, model training and testing, iterated ten times with different seeds for increased robustness. In the end, the best model is chosen to perform explainability analysis.

### A. Data description and partition

We randomly sampled ten times 400 participants from the publicly available LIDC-IDRI [4], which includes both SDCT and LDCT images. From the images, lesions $\geqslant 3$ *mm* were segmented and evaluated by one to four radiologists in terms of likelihood of malignancy on a scale of 1 (highly unlikely) to 5 (highly suspicious). The consensus segmentation was considered and multiple scores for a single nodule were averaged to derive a consensus score. In this study, nodules with an average score $\geqslant 4$ were considered likely malignant. Table I summarizes the mean and standard deviation of the number of nodules present in each randomly selected group of 400 participants across all ten seed selections. We split the 400 participants into train and test sets with an 8:2 ratio.

Accepted manuscript available at: https://doi.org/10.1109/MetroXRAINE62247.2024.10797090

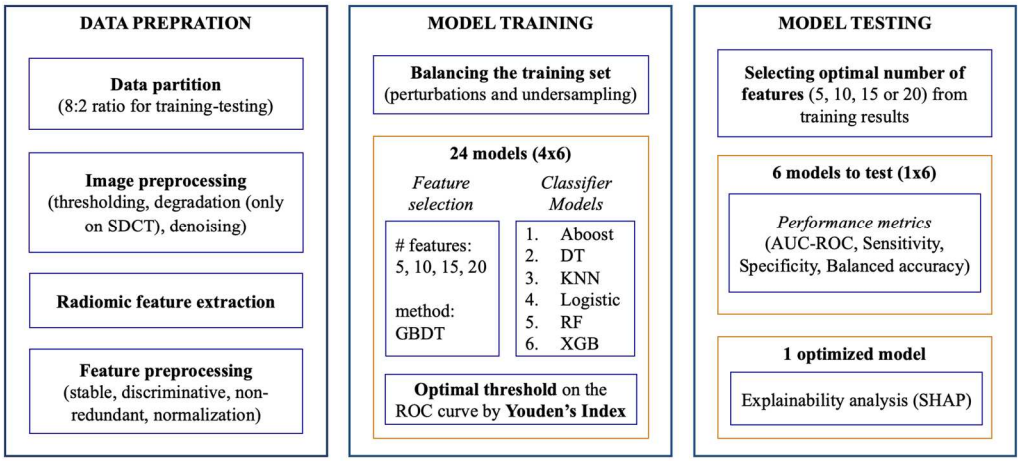


Fig. 1. Overview of the pipeline used in this study, consisted in data preparation, model training and testing. Data preparation includes partitioning, image preprocessing, radiomic feature extraction, and feature preprocessing. During model training, the training set is balanced, and the performance of six classifiers is evaluated using varying numbers of features selected with gradient boosted decision trees. The optimal ROC curve threshold is determined using Youden's index. Subsequently, models are tested with the selected optimal number of features, and the optimized model is selected based on performance metrics, followed by explainability analysis.

TABLE I. SUMMARY OF THE NUMBER OF NODULES USED IN THIS STUDY

| *Nodules* | *mean* | *std** |
|---|---|---|
| Total | 1187 | 32 |
| Train set | | |
| Likely benign | 853 | 29 |
| Likely malign | 102 | 7 |
| Test set | | |
| Likely benign | 205 | 15 |
| Likely malign | 27 | 6 |

*Standard deviation

### B. Image preprocessing

Image preprocessing involved thresholding, degradation (for SDCT only), and denoising, implemented in MATLAB version 2022b:

1. Thresholding ensured all pixels were positive by setting negative values to zero.
2. For SDCT images (CT acquired at tube current >= 80 mA), degradation based on Zeng et al.'s LDCT simulation technique [12] was applied. This method utilizes raw data from high-dose scans, obtains the raw sinogram data and injects Gaussian and Poisson noise into the sinogram domain to model electronic and quantum noise. The simulated LDCT images are reconstructed from the sinogram data. See more details in [12].
3. Denoising was performed using a 3D Gaussian filter with a $3 \times 3 \times 3$ voxel kernel and $\sigma = 0.5$.

### C. Feature extraction and preprocessing

Radiomic features were extracted from the region of interest (ROI), i.e., the segmented nodules, using Pyradiomics [13], resulting in 851 features per nodule (14 shape features, 18 first-order features, 75 texture features and 744 wavelet-based features). For more details on radiomic features, see Pyradiomics documentation [13].

Feature preprocessing was performed on the training set and applied to the test set: it comprised steps to assure feature stability, discriminant characteristics, and non-redundancy. Stability analysis assessed feature consistency under small transformations. This analysis was conducted on nodules with four radiologists' annotations using intra-class correlation coefficient (ICC). Features were deemed stable if the ICC > 0.75, as defined in [14]. Discriminative analysis identified informative features by comparing original ROIs with translated ROIs with no overlap with the nodule. In this analysis, features with ICC < 0.5 were considered discriminative. Finally, highly correlated features were removed by using the absolute Spearman's correlation coefficient > 0.85.

### D. Balancing the dataset

To address the imbalanced dataset, we first applied a data augmentation technique introduced in [15] consisting of the application of ROI perturbations on the malignant nodules of the training set (Fig. 2): 15% erosion (a) and dilation (b), as well as random contour adjustments (c). Afterwards, a random undersampling approach was adopted to reduce the non-malignant nodules to match the number of nodules in the minority class post-augmentation. Considering the ten seeds, the mean and standard deviation for nodule counts is 407 (±28) per class after balancing the dataset.

### E. Feature selection, model training and testing

Feature selection was performed on the training sets after balancing the classes using gradient-boosted decision tree (GBDT) feature selection method. The number of features selected was varied (5, 10, 15 or 20 features) centered on the sample-to-feature ratio guideline [16] suggesting "10-15 samples per feature" when the class is balanced or, when

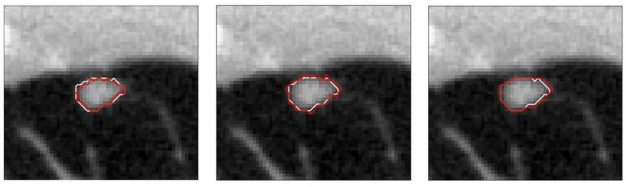


Fig. 2. Perturbation technique applied to the minority class for data augmentation: (a) erosion, (b) dilation, (c) random contour. The white contour represents the original consensus mask, while the red represents the perturbation.

Accepted manuscript available at: https://doi.org/10.1109/MetroXRAINE62247.2024.10797090

imbalanced, apply this rule to the minority class. In our case, having an average of 129 nodules in the minority class, we could select 9-13 features. Guided by this rule, we explored an interval (5-20) centered on this range (9-13). Additionally, based on our previous work [17], which extensively explored feature selection methods, supports the decision to start with 5 features and incrementally increase by 5, up to 20. Afterwards, six machine learning classifiers were trained: random forest (RF), AdaBoost (ABoost), XGBoost (XGB), decision tree (DT), K-nearest neighbors (KNN), and logistic regression (LR). The optimal cut-off on the ROC curve was determined by Youden's index during the training phase and applied to the testing.

This comprehensive analysis, resulting in a total of 24 models (4 number of features, 6 classifiers), aimed to identify the most effective number of features. To determine the optimal number of features (5, 10, 15 or 20), we compared the mean balanced accuracy of each number of features across all classifiers. After the identification of the optimal number of features, we obtain 6 models (1 number of features, 6 classifiers) which are evaluated on the test set using four performance metrics: balances accuracy, sensitivity, specificity, area under the ROC curve (AUC-ROC).

Based on these performance metrics, we selected the best-performing configuration and used SHapley Additive exPlanations (SHAP) [18] to explain the predictions.

## III. Results

A systematic analysis was conducted to assess the impact of the number of features and machine learning algorithms by computing the mean and standard deviation across multiple seeds.

### A. Optimal number of features

Fig. 3 represents the mean balanced accuracy achieved by each classifier in combination with a specific number of features selected by GBDT during the training phase. The graph shows consistent patterns of increase and decrease in mean balanced accuracy across different classifiers as the number of selected features varies. Moreover, each classifier was characterized by a general trend of increasing mean balanced accuracy as the number of selected features increased from 5 to 20 across most classifiers. Specifically, XGB consistently achieved high mean balanced accuracy scores across all feature counts. By examining the overall mean and standard deviation calculated for each number of features, it can be observed that the results for 15 features and 20 features exhibit similar mean and standard deviation (t-test: $t = 0.78$, $p > 0.05$). In a parsimony way, we decided to select 15 features as the optimal number. Table II shows the number of features selected from each category, the frequency of the features selected and the average frequency considering 10 iterations by category. For instance, there were three features (major axis length, elongation and sphericity) from the original shape category that were selected 22 times globally (i.e., major axis length was selected 10 times, elongation 7 times and sphericity 5 times). Therefore, the average frequency for a feature from this category was 7.33 (22/3) out of the 10 iterations. Comparatively, the original shape category was the most predominant category of features selected and original first-order category the least selected.

### B. Performance of classifiers in the test set

Fig. 4 shows the results on the test set of the classifiers trained with 15 features. The values reported were computed after utilizing the optimal cut-off determined by Youden's index during the training phase.

In terms of balanced accuracy and AUC-ROC, RF, KNN and LR demonstrated comparable performance, overperforming the other classifiers. When examining sensitivity and specificity, a trade-off is particularly visible in ABoost and XGB classifiers, where high specificity values are associated with low sensitivity. In contrast, RF classifier maintained high scores for both sensitivity and specificity compared to the other classifiers.

TABLE II. Frequency and Average Selection of Features by Category over 10 Iterations.

| *Feature category* | | *Nr. features* | *Total times selected* | *Average* |
|---|---|---|---|---|
| Original | Shape | 3 | 22 | 7.33 |
| | First-order | 2 | 2 | 1 |
| | Textural | 6 | 10 | 1.67 |
| Wavelet | First-order | 23 | 43 | 1.87 |
| | Textural | 43 | 73 | 1.70 |

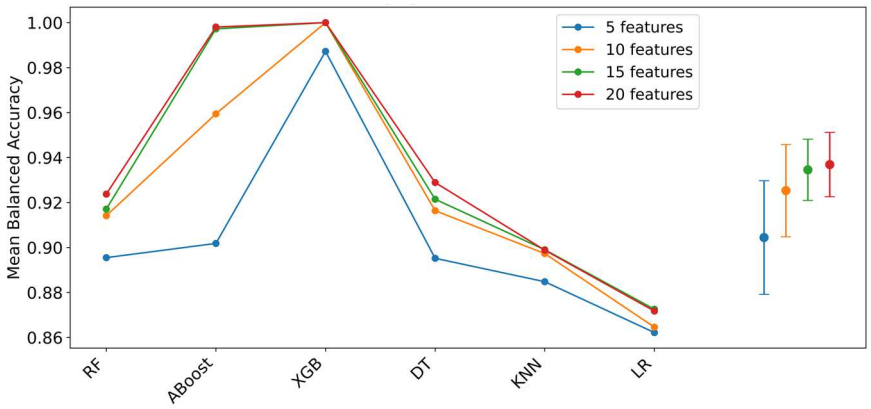


Fig. 3. Mean balanced accuracy based on the 10 seeds achieved by each classifier in combination with a specific number of features (5, 10, 15 and 20) selected by GBDT during the training phase.

Accepted manuscript available at: https://doi.org/10.1109/MetroXRAINE62247.2024.10797090

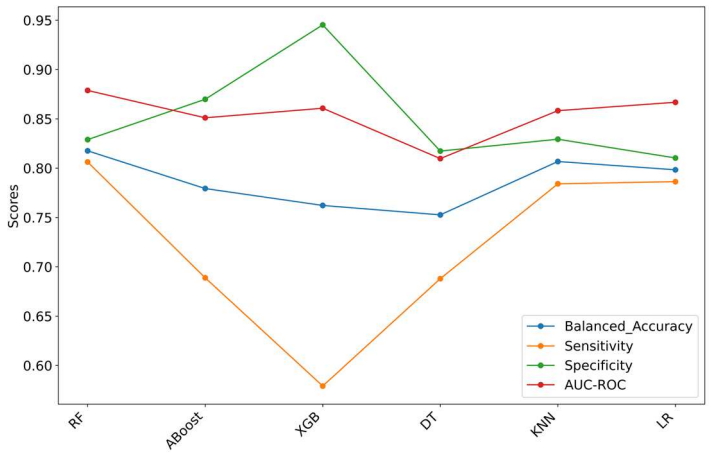


Fig. 4. Results achieved by different classifiers using 15 features selected by gradient boost decition tree on the test set.

Considering all performance metrics, RF emerged as the best performing classifier, achieving for balanced accuracy, sensitivity, specificity and AUC-ROC scores of 0.82 ± 0.04, 0.81 ± 0.08, 0.83 ± 0.05, 0.88 ± 0.05, respectively.

### C. *Explainability analysis*

Considering RF model for explainability analysis, the top three most important features identified from the SHAP analysis (Fig. 5) were: *MajorAxisLength*, *sphericity* and *wavelet_LHH_ngtdm_Busyness*. The feature *MajorAxisLength*, representing the largest axis length of the ROI-enclosing ellipsoid, showed a positive correlation with malignancy likelihood, indicating that medium to higher values are associated with a greater likelihood of malignancy. Conversely, *sphericity*, which measures the roundness of the shape of the tumor region relative to a sphere, demonstrated an inverse relationship with malignancy likelihood. Specifically, lower values of sphericity (more irregular the nodule) were found to be associated with a higher likelihood of malignancy. The *wavelet_LHH_ngtdm_Busyness* feature is a textural feature computed from neighboring gray tone difference matrix (NGTDM) obtained after applying low-pass (L) and high-pass (H) filters. It measures the intensity change of a pixel considering its neighbors. A low busyness value indicates a gradual change of intensity between pixels and their neighborhood, which was overall associated with a lower likelihood of malignancy.

### D. *Comparison with original malignancy scores*

As the original malignancy scores given by the radiologists ranged from 1 to 5, with nodules having an average score $\geqslant$ 4 considered malignant, we sought to determine how classified and misclassified instances in the test set were spread across malignancy levels. This analysis was performed for the RF model using the 15 features selected by GBDT. The histogram in Fig. 6 reveals that a significant portion of misclassified nodules belonged to the category with an average score between 3 and 4, representing the most uncertain cases. In contrast, none of the nodules from class [1;2[ (average score less than 2) were misclassified and there were very few misclassifications from the class [4;5] (average score of 4 or higher).

## IV. Discussion

In the present study, a novel explainable radiomic-based pipeline was developed using degraded SDCT and LDCT images to classify lung nodules based on the likelihood of malignancy. A key innovation stands in degrading SDCT images to leverage the annotations present in the LIDC-IDRI dataset, aiming to obtain a screening scenario where only LDCTs are present while addressing the limitations associated with different dosage CT images in the same dataset. Furthermore, we also addressed the challenge of imbalanced datasets by implementing a recent data augmentation technique involving creating perturbations in nodules present in the minority class and random undersampling of the majority class. To ensure the robustness of our findings, we conducted the pipeline ten times with varying seeds, enabling the generation of multiple data partitions. This approach enhances the reliability of our testing results by mitigating the influence of seed selection bias.

The final optimized model, consisting of 15 features selected by GBDT and trained with RF classifier, achieved the highest performance scores with balanced accuracy of

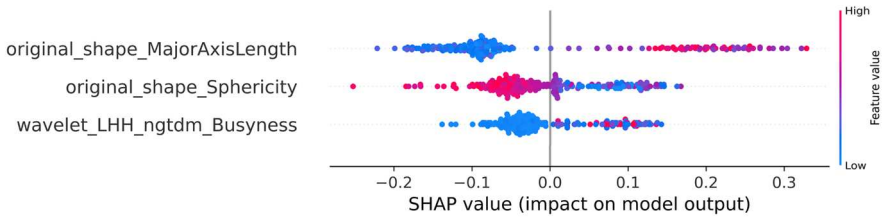


Fig. 5. SHAP beeswarm plot of the test set (seed 8) generated from the optimized model, 15 features selected by GBDT with RF classifier.

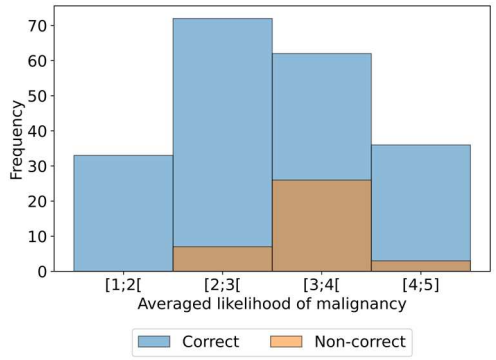


Fig. 6. Histogram representing the original label distribution (average likelihood of malignancy) of the correctly and incorrectly classified instances in the test set (seed 8).

0.82 ± 0.04, sensitivity of 0.81 ± 0.08, specificity of 0.83 ± 0.05 and auc-roc of 0.88 ± 0.05. These results are comparable to or exceeded the performance of several studies in the literature (Table III), such as Alahmari et al. [9] (sensitivity of 0.49 and specificity of 0.93) and Rundo et al. [10] (sensitivity of 67.0% and specificity of 86.2%). Causey et al. [8] achieved notable results (sensitivity of 0.88 and specificity of 0.88) when using deep learning convolutional neural network features on the LIDC-IDRI dataset. However, when applying only radiomic features the sensitivity dropped to 0.70. Moreover, Causey et al. [8] also omitted nodules with an average score for the likelihood of malignancy between 3-4, which are the most challenging nodules to classify.

We recognize that larger datasets contribute to more stable and accurate models, enhancing reliability and generalizability. Our dataset comprised a larger number of nodules (400), surpassing the sample sizes of studies that achieved better performance with smaller datasets. For example, Choi et al. [6] considered 72 subjects, Mao et al. [19] 295 nodules, Garau et al. [7] 110 subjects, and Liu et al. [20] 141 nodules.

Although Peikert et al. [11] achieved better performance metrics (sensitivity = 0.90, specificity = 0.86) with a relatively large dataset, it is worth noting that their study included nodules larger than those analyzed in our study. Their dataset comprised nodules with a size threshold of ⩾7 *mm*, whereas our study focused on nodules with a size threshold of ⩾3 *mm*, considered optimal for screening [21]. In addition, research suggests that the lower the size of the nodule, the higher the rate of false positives [21], emphasizing the need for robust classification models in this context. Furthermore, Peikert et al. [11] selected data to balance the dataset, potentially introducing bias.

TABLE III. COMPARISON WITH RESULTS FOUND IN THE LITERATURE.

| *Author* | *N* | *Nodule size (mm)* | *Sen. %* | *Spec.%* | *AUC* |
|---|---|---|---|---|---|
| Choi et al. [6] | 72 | ⩾ 3 | 87.2 | 81.2 | 0.89 |
| Peikert et al. [11] | 726 | ⩾ 7 | 90.4 | 85.5 | 0.94 |
| Alahmarie et al.[9] | 467 | ⩾ 4 | 49.0 | 93.0 | 0.82 |
| Causey et al.[8] | 664 | ⩾ 3 | 69.7 | 95.5 | 0.94 |
| Mao et al. [19] | 294 | ⩾ 6 | 81.0 | 92.2 | 0.90 |
| Garau et al. [7] | 110 | - | 90.0[a], 35.5[b] | | 0.86 |
| Liu et al. [20] | 141 | - | 86.4 | 85.7 | 0.89 |
| Rundo et al. [10] | 11 012 | - | 67.0 | 86.2 | 0.82 |
| **Proposed** | **400** | **⩾ 3** | **81.0** | **83.0** | **0.88** |

N: sample size; [a] true positive rate; [b] false positive rate.

Additionally, we employed the SHAP method to interpret the model's classification and identify key features. This interpretability analysis enabled us to mimic radiologists' scoring processes, providing valuable insights into the decision-making process. Specifically, identifying *MajorAxisLength* and *sphericity* as the most important features aligns with criteria used in the Lung Imaging Reporting and Data System (Lung-RADS) developed by the American College of Radiology to standardize the screening of lung cancer on CT images [22]. Moreover, the inclusion of textural wavelet features (*wavelet_LHH_ngtdm_Busyness)* among the top three important features suggests that the wavelet transform provided additional information not visible to the naked eye.

The present study is not exempt from limitations. For instance, the dataset under consideration was primarily curated to develop segmentation methods rather than for classification purposes or screening. In addition, nodule evaluation relied on radiologist assessments rather than histology-confirmed malignancy, leading to subjectivity in nodule evaluation.

However, our study successfully emulated the radiologists' approach to nodule evaluation, demonstrating the effectiveness of combining visually perceptible shape and size features with wavelet features, which offer insights beyond human perception.

## V. CONCLUSION

In conclusion, the approach proposed herein demonstrated the potentiality of utilizing degrading SDCT images as a promising method for overcoming challenges associated with limited annotated LDCT images, as well as radiomics in the discrimination between the likely malign nodules and less likely ones.

## REFERENCES

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] National Lung Screening Trial Research Team, "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, Aug. 2011, doi: 10.1056/NEJMoa1102873.

[3] A. Khawaja *et al.*, "Do we need to see to believe?—radiomics for lung nodule classification and lung cancer risk stratification," *J Thorac Dis*, vol. 12, no. 6, p. 3303, Jun. 2020, doi: 10.21037/JTD.2020.03.105.

[4] S. G. Armato *et al.*, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans," *Med Phys*, vol. 38, no. 2, p. 915, 2011, doi: 10.1118/1.3528204.

[5] F. Zhu, C. Yang, J. Zou, W. Ma, Y. Wei, and Z. Zhao, "The classification of benign and malignant lung nodules based on CT radiomics: a systematic review, quality score assessment, and meta-analysis," *https://doi.org/10.1177/02841851231205737*, vol.

64, no. 12, pp. 3074–3084, Oct. 2023, doi: 10.1177/02841851231205737.

[6] W. Choi *et al.*, "Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer," *Med Phys*, vol. 45, no. 4, pp. 1537–1549, Apr. 2018, doi: 10.1002/mp.12820.

[7] N. Garau *et al.*, "External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis," *Med Phys*, vol. 47, no. 9, pp. 4125–4136, Sep. 2020, doi: 10.1002/mp.14308.

[8] J. L. Causey *et al.*, "Highly accurate model for prediction of lung nodule malignancy with CT scans OPEN", doi: 10.1038/s41598-018-27569-w.

[9] S. S. Alahmari, D. Cherezov, D. B. Goldgof, L. O. Hall, R. J. Gillies, and M. B. Schabath, "Delta Radiomics Improves Pulmonary Nodule Malignancy Prediction in Lung Cancer Screening," *IEEE Access*, vol. 6, pp. 77796–77806, 2018, doi: 10.1109/ACCESS.2018.2884126.

[10] L. Rundo *et al.*, "A low-dose CT-based radiomic model to improve characterization and screening recall intervals of indeterminate prevalent pulmonary nodules," *Diagnostics*, vol. 11, no. 9, p. 1610, Sep. 2021, doi: 10.3390/diagnostics11091610.

[11] T. Peikert *et al.*, "Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the National Lung Screening Trial," *PLoS One*, vol. 13, no. 5, May 2018, doi: 10.1371/JOURNAL.PONE.0196910.

[12] D. Zeng *et al.*, "A Simple Low-dose X-ray CT Simulation from High-dose Scan," *IEEE Trans Nucl Sci*, vol. 62, no. 5, p. 2226, Oct. 2015, doi: 10.1109/TNS.2015.2467219.

[13] J. J. M. Van Griethuysen *et al.*, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Res*, vol. 77, no. 21, p. e104, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[14] T. K. Koo and M. Y. Li, "Cracking the Code: Providing Insight Into the Fundamentals of Research and Evidence-Based Practice A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," 2016, doi: 10.1016/j.jcm.2016.02.012.

[15] F. Lo Iacono, R. Maragna, G. Pontone, and V. D. A. Corino, "A Novel Data Augmentation Method for Radiomics Analysis Using Image Perturbations," *Journal of Imaging Informatics in Medicine 2024*, pp. 1–14, May 2024, doi: 10.1007/S10278-024-01013-0.

[16] J. D. Shur *et al.*, "Radiomics in oncology: A practical guide," *Radiographics*, vol. 41, no. 6, pp. 1717–1732, Oct. 2021, doi: 10.1148/rg.2021210037.

[17] J. Liu *et al.*, "Developing a robust two-step machine learning multiclassification pipeline to predict primary site in head and neck carcinoma from lymph nodes," *Heliyon*, vol. 10, no. 2, Jan. 2024, doi: 10.1016/J.HELIYON.2024.E24377.

[18] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Conference on Neural Information Processing Systems*, 2017, Accessed: Mar. 07, 2024. [Online]. Available: https://github.com/slundberg/shap

[19] L. Mao *et al.*, "Quantitative radiomic model for predicting malignancy of small solid pulmonary nodules detected by low-dose CT screening," *Quant Imaging Med Surg*, vol. 9, no. 2, pp. 263–272, 2019, doi: 10.21037/qims.2019.02.02.

[20] J. Liu *et al.*, "Comparison of Radiomic Models Based on Low-Dose and Standard-Dose CT for Prediction of Adenocarcinomas and Benign Lesions in Solid Pulmonary Nodules," *Front Oncol*, vol. 10, p. 1, Feb. 2021, doi: 10.3389/fonc.2020.634298.

[21] M. Silva, G. Milanese, R. E. Ledda, U. Pastorino, and N. Sverzellati, "Screen-detected solid nodules: from detection of nodule to structured reporting," *Transl Lung Cancer Res*, vol. 10, no. 5, p. 2335, May 2021, doi: 10.21037/TLCR-20-296.

[22] American College of Radiology Committee on Lung-RADS, "Lung-RADS Assessment Categories 2022." Accessed: May 08, 2024. [Online]. Available: https://www.acr.org/-/media/ACR/Files/RADS/Lung-RADS/Lung-RADS-2022.pdf