2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) | 979-8-3503-9412-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/SYNASC61333.2023.00040

# Prediction of Malignancy in Lung Cancer using several strategies for the fusion of Multi-Channel Pyradiomics Images

1st Guillermo Torres
*Computer Science Dept., UAB*
*Computer Vision Center*
Barcelona, Spainy
gtorres@cvc.uab.cat

2nd Jan Rodriguez Dueñas
*Universitat Autònoma de Barcelona (UAB)*
Barcelona, Spain
jan.rodriguez@autonoma.cat

3rd Sonia Baeza Mena
*Respiratory Medicine Dept.*
*H. U. Germans Trias i Pujol (HUGTP)*
Barcelona, Spain
smbaeza.germanstrias@gencat.cat

4th Antoni Rosell Gratacós
*Respiratory Medicine Dept.*
*H. U. Germans Trias i Pujol (HUGTP)*
Barcelona, Spain
arosellg.germanstrias@gencat.cat

5th Carles Sanchez
*Computer Science Dept., UAB*
*Computer Vision Center*
Barcelona, Spainy
csanchez@cvc.uab.cat

6th Debora Gil
*Computer Science Dept., UAB*
*Computer Vision Center*
Barcelona, Spain
debora@cvc.uab.cat

***Abstract*—Radiomics has become a reference tool for the early diagnosis of cancer in screening programs. Radiomics involves the extraction of a large number of quantitative features from medical scans. Such features define a representation space of lesions that should capture the heterogeneity of tissue texture in order to discriminate malignacy.**

**This work presents the process of generation and subsequent study of a representation space of pulmonary nodules given by the deep convolutional features of multiple texture images extracted from computed tomography (CT). The objective of the study is to assess whether this deep texture features have a positive impact on the diagnosis of lung cancer compared to deep intensity features. To do so, we have trained SVM models with different data splits, evaluating the diagnostic performance using metrics defined at, both, slice and nodule levels on an own data base for early diagnosis of small (<3cm) pulmonary nodule. Results show that deep texture features perform better than intensity ones with a raise in specificity from 0.49 to 1 at slice level and 0 to 0.67 at nodule level.**

***Index Terms*—Lung cancer screening, early diagnosis, radiomics, representation space**

## I. Introduction

Early detection of lung cancer (LC) plays a crucial role in improving treatment outcomes and patient prognosis. Studies such as the National Lung Screening Trial (NLST) [1] and NELSON [2] have demonstrated that annual screening with low-dose computed tomography (LDCT) can effectively reduce mortality rates associated with LC [3]. Approximately 12-13% of LDCT scans yield positive results for the detection of pulmonary nodules (PNs). Among these detected PNs, around 60% necessitate follow-up with additional imaging, and approximately 40% of them (which accounts for 5% of the total LDCTs performed) require more intensive monitoring and closer follow-up.

Radiomics involves the extraction of a large number of quantitative features from medical images, such as computed tomography (CT) scans, magnetic resonance imaging (MRI), or positron emission tomography (PET). These features capture the heterogeneity and characteristics of lung tumors at a microscopic level, enabling more precise diagnosis and treatment planning. In a pilot study [4], image features extracted from NLST (National Lung Screening Trial) data using radiomic techniques exhibited superior predictive value compared to volumetric measurements alone and, later, the authors in [5] developed a radiomic classifier incorporating location variables, size, shape descriptors, and texture analysis.

Machine learning approaches [6] based on features like Gabor, Local Binary Patterns (LBP), and SIFT descriptor in combination with classifiers such as Support Vector Machine (SVM) and Random Forest, have shown improved diagnostic power with high sensitivity and specificity, achieving an AUC of 0.97 and sensitivity of 96% with 95% specificity.

Recently, and motivated by its performance in other areas of application, researchers have began to classify PN by using CNNs. The early work of Shen et al. proposed to use a multi-crop CNN [7] to make the model robust to scales of nodules while keeping 2D input images. Results showed an overall accuracy (including malign and benign cases) of 87%. However, the authors did not report sensitivity for malignancy detection and specificity for discarding benign nodules and, thus, its true clinical value is uncertain.

Since nodules are 3D structures, some works have addressed the problem using 3D CNNs. Yan et al. [8] explored 3D CNNs for pulmonary nodule classification in comparison to a slice-level 2D CNN and a nodule-level 2D CNN analysis. The 3D approach was the best performer with a 87% of overall accuracy and similar specificity and sensitivity at the cost of

979-8-3503-9412-2/23/$31.00 ©2023 IEEE
DOI 10.1109/SYNASC61333.2023.00040

a significantly higher demand of computational resources and annotated data. Zhu et al. [9] used 3D deep dual path networks (DPNs) a 3D Faster Regions with Convolutional Neural Net (R-CNN) designed for nodule detection with 3D dual path blocks and a U-net-like encoder-decoder structure to effectively learn nodule features. Despite the complex architecture used, this approach could only achieve a 81% of sensitivity and specificity was not reported. Jiang et al. [10] sequentially deployed a contextual attention module and a spatial attention module to 3D DPN to increase the representation ability. A main novelty of this work is that it ensembles different model variants to improve the prediction robustness. Results show an increase of sensitivity to 90% while keeping a specificity similar to [8].

GLCM (Gray-Level Co-occurrence Matrix) texture features, have demonstrated effectiveness in cancer diagnosis across various medical imaging modalities [11]. In a recent study [12], researchers proposed a hybrid approach that combined GLCM textural features with a neural network for nodule characterization in CT scans. To ensure reproducibility with limited training data, an embedding technique based on the statistical significance of radiomic features was used. This embedded representation served as the input for a neural network, with its architecture and hyperparameters optimized using custom-defined metrics. The best performing model achieved a sensitivity of 100% and specificity of 83% (with an AUC of 0.94) for malignancy detection when evaluated on an independent patient set. This innovative approach shows promise in improving the accuracy and reliability of lung cancer screening by integrating radiomic features and deep learning techniques, offering potential solutions to the challenges posed by false positives in current screening methods.

In this work we present a strategy for malignancy detection based on deep textural features extracted using texture images and VGG16 (Section II), as well as, the validation protocol for the assessment of its level of generalization (Section II-A).

## II. Strategy for Diagnosis of Malignancy

Our workflow consists of multiple steps, which are illustrated in Figure 1. First, we extract the nodule region of interest (ROI) from CT scans using a predefined ROI. Subsequently, in the generation of the representation space defining a visual embedding of the nodule, we have the option to either pass the ROI without any modifications or extract GLCM features from it. These features are then fed into a pre-trained VGG16 network to obtain the final feature embeddings. We have explored three different strategies for feature fusion to combine these embeddings and train a model for predicting the axial 2D images of the nodule ROI. The final nodule diagnosis is determined using a max-voting criteria.

The nodule ROIs are defined by a radiologist using the software 3D-Slicer, which is a free, open-source software for visualization and processing medical images. The ROI always includes the intranodule region (the nodule itself), and the perinodular region (the area around the nodule). Studies like [13], [14] have emphasized the significance of incorporating the perinodular region for precise classification of benign and malignant nodules.

We use two methods, GLCM Textural Features and Gray Level-Intensity, to generate the representation space. These methods utilize the extracted nodule ROIs, as shown in Figure 1. The goal of this feature embedding step is to derive meaningful and discriminative representations of the nodules, which can be further analyzed and used for classification tasks.

The GLCM textural features are calculated using the nodule ROI. Additionally, for each nodule, we generate a fictitious nodule mask where all voxel values are set to one. This mask indicates that all voxels within the nodule ROI should be considered when computing the GLCM features. By employing this nodule mask, we can generate 21 GLCM features (i.e., 21 volumes) for each nodule ROI, corresponding to the textural features computed in [12]. GLCM features are statistical descriptors computed from a gray-level co-occurrence matrix. This matrix captures the frequency of occurrence of pixel pairs with specific gray-level values and spatial relationships within a defined neighborhood.

To generate the GLCM features, image intensity is discretized using the histogram of the original volume intensity into $n$ discrete bins. The width of these histogram bins determines the level of granularity at which the GLCM features describe the textural patterns. Smaller bin widths provide a finer level of detail, while larger bin widths result in more generalized information. Once the gray values are discretized, the GLCM is constructed by examining the spatial relationships between pixels within the neighborhood. Specifically, for each pixel, the occurrence of gray-level pairs and their spatial relationships with neighboring pixels are recorded in the co-occurrence matrix. Based on the GLCM, a variety of statistical measures (including contrast, correlation, energy, homogeneity, and many others) are computed to extract textural information.

To extract deep features, we use the axial slices of the original intensity volumes and the 21 GLCM textural volumes as inputs, which are then passed through a pre-trained VGG16 model that was externally trained on ImageNet [15]. The VGG16 architecture is composed of 13 convolutional layers, 5 max-pooling layers (2 × 2), and 3 fully-connected layers (namely FC6, FC7 and FC8). The linear output layer utilizes the softmax activation function. ReLU activation function is applied to all the convolutional layers, while dropout regularization is employed in the fully connected layers. The deep representation for both intensity and texture images is derived from the features extracted from the FC6 layer, the first fully connected layer in the VGG16 model, situated after the convolutional layers.

For each image, the deep feature vector from the FC6 layer has a dimensional size of 4096. In the case of intensity images, this results in 4096 features. However, for the GLCM features approach, which includes 21 GLCM volumes per nodule, the resulting features have a dimension of (21, 4096). These 21 channels need to be combined to create an input vector for a classifier. There are three options considered: concatenation,

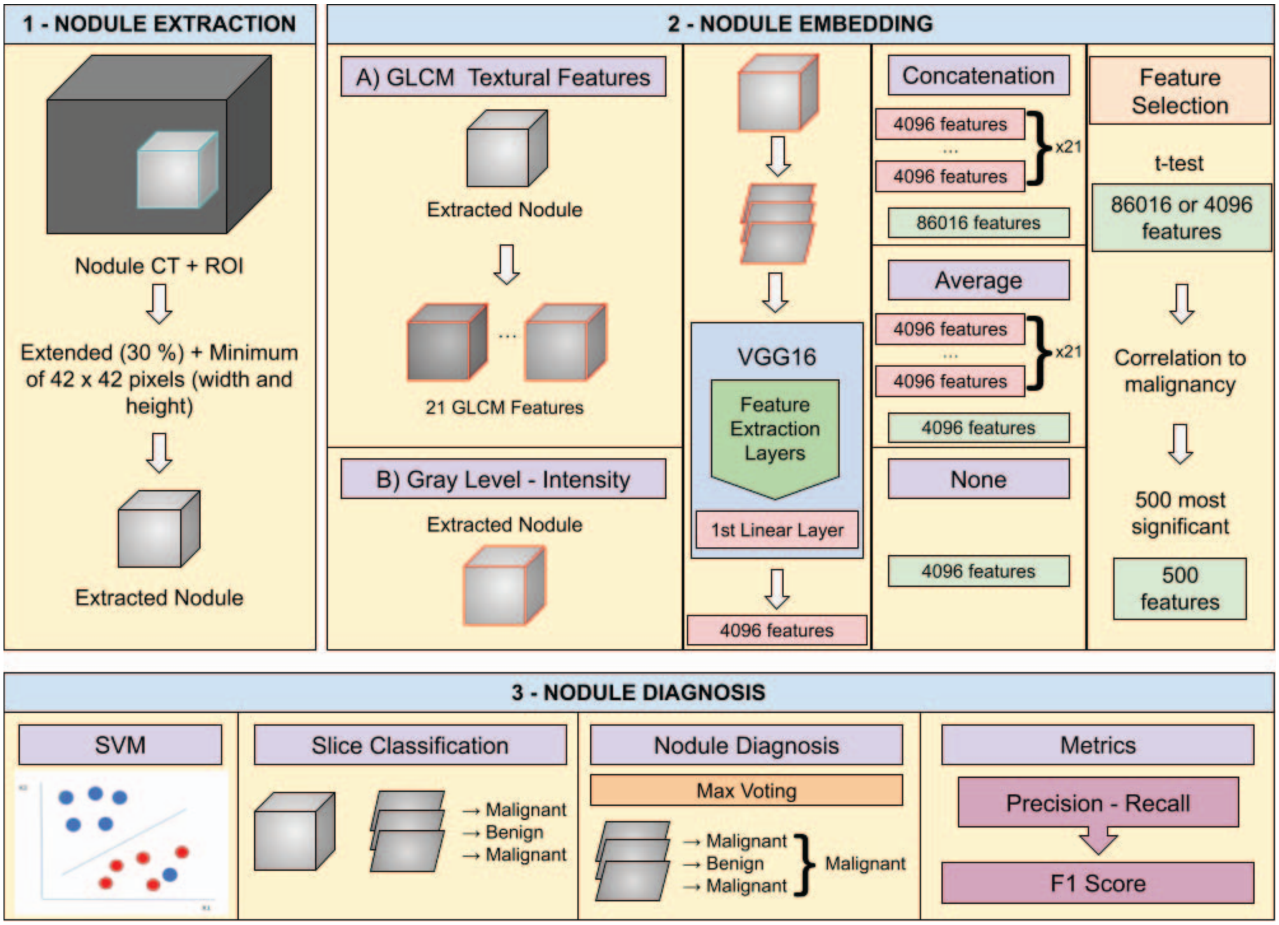

Fig. 1. Workflow of the strategy for diagnosis of malignancy.

average, and none.

In the concatenation strategy, the features are flattened, resulting in 86016 features. This means that the 21 channels are concatenated to form a single long feature vector. In the average strategy, the features are used to compute an average. This results in a single feature vector with the same dimension as each individual feature vector, i.e., (1, 4096). Lastly, for the raw gray levels features, they result in features with a dimension of (1, 4096), meaning that there is only one channel in the feature vector.

Regardless of the chosen strategy for features fusion, we proceed to apply a t-test to rank the features based on their significance in correlating with nodule malignancy. This step enables us to perform feature selection and identify the most relevant features. For the VGG features, they are ranked based on the p-value obtained from a t-test that measures the difference in averages between malignant and benign slices. The top 500 features with the lowest p-values are then selected as input for the SVM classifier.

The fused features (none, concatenation, and average) are used to train an SVM classifier for making slice-by-slice predictions. In order to optimize the SVM parameters, we perform a grid search method where multiple combinations of the parameters C, kernel, and gamma are tested. After training the SVM, the nodule diagnosis is determined by max-voting of the slice predictions.

### A. Levels of Generalization

We split the data into three different levels of generalization to study the impact of the new representation space. Each level uses a distinct experimental sampling unit (either nodule or slice) to split data into training and test folds.

1) **Nodule k-folds.** In this approach, the sampling unit for data splitting is the nodule, this way the k-fold cross-validation assesses the performance of our model when predicting unseen data. This is consequence of having slices in the test set that are completely independent from the ones used in training, as they belong to different nodules. The capability for reproducing results in new unseen data can be assessed by computing statistical ranges from metrics computed for the test folds.
2) **Leave-1-Nodule-Out** This approach represents a particular implementation of a k-fold cross-validation, with k set to the maximum number of nodules in the dataset.

The nodule serves as the experimental unit for data splitting, but unlike a k-fold splitting only average metrics can be computed. Therefore the capability for generalising can not be statistically assessed.

3) **Slice k-folds.** In this approach (followed by the majority of SoA methods), the sampling unit for data splitting is the nodule slices without stratifing by nodule. This implies that slices from the same nodule can be present in, both, the training and test data. Therefore the test set is not independent from the one used for training and, consequently, the statistical ranges computed accross test folds are over-optimistic and do not ensure reproducibility of results on new unseen data.

## III. Experiments

This study utilizes our database[1], which comprises patients recruited from the Germans Trias i Pujol University Hospital (HUGTiP) in Barcelona, Spain. The database includes images and clinical/demographic data collected between December 2019 and November 2022. A total of 92 patients with a single PN with diameter between 8 to 30 mm were included in this study. All patients underwent low-dose CT-chest scans and had pulmonary nodules (PN) that required surgical intervention.

The minimum size of a nodule region of interest (ROI) is 42x42 pixels (width and height), as imposed by the pretrained VGG16 network. For the computation of the 21 GLCM features [16], we used PyRadiomics [17] (version 3.01) with a ($3 \times 3 \times 3$) kernel and the image values discretized into 128 bins. The number of features selected by the t-test is 500.

Regarding data spliting described in Section II-A, we have used 5 folds for, both, spliting at nodule and slice levels using the python StratifiedGroupKFold function to ensure the same proportion of classes in, both, train and test. Besides, for the slice split, 25 ($\approx$ 30%) nodules of the dataset were randomly selected as a independent set (Holdout) of test patients to assess the reproducibility of the ranges computed in a slice split. We have used precision, recall and the F1-score as quality metrics.

The results obtained for the optimal configurations are summarized (mean ± standard deviation) in Table I. We notice that the Intensity domain has the lowest score among all domains. When using slice folds for splitting, both GLCM-Concatenation and GLCM-Average domains exhibit high recall for both benign and malignant nodules. The recall range for GLCM-Concatenation is (1, 1) for malignant cases and (0.84, 1) for benign cases. However, when splitting at the nodule level, the GLCM-Average domain experiences a significant drop in benign recall, almost reaching 0. On the other hand, for the GLCM-Concatenation domain, while the malignancy recall score falls within the range of (0.88, 1), the recall range for benign cases is (0.37, 1). It is worth noting that the high standard deviation (around 30%) indicates considerable variability across folds for the GLCM-Concatenation domain. This variability is attributed to the limited number of benign samples (3 at most) of the test set.

## IV. Conclusions

The domain with the lowest performance is the Intensity, which can be attributed to the fact that VGG16 was trained on ImageNet and, thus, may not effectively capture the texture details characteristic of cancer tumor lesions. On the other hand, GLCM demonstrates higher discrimination power as it can represent the texture details of the nodules. Regardless of the representation space, the data split at the slice level yields the least reproducible results with over optimistic metrics. In particular we achieve average sensitivity of 0.98 with 1.0 of specificity, which are comparable to those achieved by state-of-the-art methods. The interval predictions obtained by splitting the data at the nodule level are less optimistic but more realistic as they include the hold-out metric results.

These observations highlight the challenges and limitations in achieving consistent and reliable results in lung cancer screening. The findings underscore the need for further research and development to address issues related to dataset size, imbalance and reproducibility, ultimately improving the accuracy and reliability of screening methods.

## Acknowledgment

This project is supported by the Ministerio de Economía, Industria y Competitividad, Gobierno de España grant number PID 2021-126776OB-C21, Agència de Gestió d'Ajuts Universitaris i de Recerca grant numbers 2021SGR01623 and CERCA Programme / Generalitat de Catalunya.

## References

[1] N. L. S. T. R. Team, "The national lung screening trial: overview and study design," *Radiology*, vol. 258, no. 1, pp. 243–253, 2011.

[2] H. J. de Koning, C. M. van der Aalst, P. A. de Jong *et al.*, "Reduced lung-cancer mortality with volume ct screening in a randomized trial," *New England journal of medicine*, vol. 382, no. 6, pp. 503–513, 2020.

[3] C. for Disease Control and Prevention, "Who should be screened for lung cancer?" 2022, accessed June 29, 2023. [Online]. Available: https://www.cdc.gov/cancer/lung/basic_info/screening.htm

[4] Y. Liu, J. Kim, Y. Balagurunathan *et al.*, "Prediction of pathological nodal involvement by ct-based radiomic features of the primary tumor in patients with clinically node-negative peripheral lung adenocarcinomas," *Medical physics*, vol. 45, no. 6, pp. 2518–2526, 2018.

[5] T. Peikert, F. Duan, S. Rajagopalan, R. A. Karwoski *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, p. e0196910, 2018.

[6] F. Zhang, Y. Song, W. Cai *et al.*, "Lung nodule classification with multilevel patch-based context analysis," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 4, pp. 1155–1166, 2013.

[7] W. Shen, M. Zhou, F. Yang *et al.*, "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification," *Pattern Recognition*, vol. 61, pp. 663–673, 2017.

[8] X. Yan, J. Pang, H. Qi *et al.*, "Classification of lung nodule malignancy risk on computed tomography images using convolutional neural network: A comparison between 2d and 3d strategies," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 91–101.

[9] W. Zhu, C. Liu, W. Fan, and X. Xie, "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 673–681.

[1] The URL to download our database will be made available in the camera's ready stage.

TABLE I
DIAGNOSIS SCORE AT NODULE LEVEL IN EXPERIMENTS WITH SVM

| Data Domain | Split | Diagnosis | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Intensity | Nodule 5-folds | Malign | 0.85 ± 0.04 | 1.00 ± 0.00 | 0.92 ± 0.02 |
| | | Benign | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 |
| | L10 | Malign | 0.86 | 1.00 | 0.92 |
| | | Benign | 1.00 | 0.07 | 0.13 |
| | Slice 5-folds | Malign | 0.95 ± 0.01 | 1 ± 0.00 | 0.98 ± 0.01 |
| | | Benign | 1 ± 0.00 | 0.49 ± 0.09 | 0.65 ± 0.08 |
| | Holdout | Malign | 0.69 | 1.00 | 0.82 |
| | | Benign | 0.00 | 0.00 | 0.00 |
| GLCM-Concatenation | Nodule 5-folds | Malign | 0.94 ± 0.05 | 0.94 ± 0.06 | 0.94 ± 0.04 |
| | | Benign | 0.70 ± 0.27 | 0.67 ± 0.30 | 0.63 ± 0.22 |
| | L10 | Malign | 0.90 | 0.95 | 0.93 |
| | | Benign | 0.56 | 0.39 | 0.46 |
| | Slice 5-folds | Malign | 1 ± 0.00 | 0.99 ± 0.01 | 0.99 ± 0.00 |
| | | Benign | 0.91 ± 0.07 | 1 ± 0.00 | 0.95 ± 0.04 |
| | Holdout | Malign | 0.75 | 0.83 | 0.79 |
| | | Benign | 0.40 | 0.29 | 0.33 |
| GLCM-Average | Nodule 5-folds | Malign | 0.87 ± 0.04 | 1.00 ± 0.0 | 0.93 ± 0.02 |
| | | Benign | 0.20 ± 0.40 | 0.10 ± 0.20 | 0.13 ± 0.27 |
| | L10 | Malign | 0.86 | 1.00 | 0.92 |
| | | Benign | 0.00 | 0.00 | 0.00 |
| | Slice 5-folds | Malign | 0.99 ± 0.01 | 1 ± 0.0 | 1 ± 0.00 |
| | | Benign | 1 ± 0.0 | 0.93 ± 0.09 | 0.96 ± 0.05 |
| | Holdout | Malign | 0.72 | 1.00 | 0.84 |
| | | Benign | 1.00 | 0.14 | 0.25 |

[10] H. Jiang, F. Gao, X. Xu, F. Huang, and S. Zhu, "Attentive and ensemble 3d dual path networks for pulmonary nodules classification," *Neurocomputing*, vol. 398, pp. 422–430, 2020.

[11] C.-L. Huang, M.-J. Lian, Y.-H. Wu, W.-M. Chen, and W.-T. Chiu, "Identification of human ovarian adenocarcinoma cells with cisplatin-resistance by feature extraction of gray level co-occurrence matrix using optical images," *Diagnostics*, vol. 10, no. 6, p. 389, 2020.

[12] G. Torres, S. Baeza, C. Sanchez *et al.*, "An intelligent radiomic approach for lung cancer screening," *Applied Sciences*, vol. 12, no. 3, p. 1568, 2022.

[13] N. Beig, M. Khorrami, M. Alilou, P. Prasanna, N. Braman, M. Orooji, S. Rakshit, K. Bera, P. Rajiah, J. Ginsberg *et al.*, "Perinodular and intranodular radiomic features on lung ct images distinguish adenocarcinomas from granulomas," *Radiology*, vol. 290, no. 3, pp. 783–792, 2019.

[14] J. L. L. Calheiros, L. B. V. de Amorim, L. L. de Lima, A. F. de Lima Filho, J. R. Ferreira Júnior, and M. C. de Oliveira, "The effects of perinodular features on solid lung nodule classification," *Journal of Digital Imaging*, pp. 1–13, 2021.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[17] AIM-Harvard, "Pyradiomics: an open-source python package for the extraction of radiomics features from medical imaging." 2016, accessed March 29, 2023. [Online]. Available: https://pyradiomics.readthedocs.io/en/latest/