Performance of Radiomics Based Multi-Parameter Models Derived from Lung CT in Differentiating Benign Non-Inflammatory, Inflammatory and Malignant Pulmonary Nodule

Angelina Tseng

Havergal College, North York, ON, Canada

angelinatseng888@gmail.com

Abstract. Pulmonary nodules are a significant clinical issue that require accurate and efficient diagnosis. This study constructed a machine learning model, combining radiomics features of chest CT sequences with three types of microvascular density (MVD) values, to differentiate benign, inflammatory and malignant pulmonary nodules. A total of 100 patients with lung nodules on CT images and corresponding pathological results were retrospectively included in the study. The MVD values and radiomics features were calculated and extracted based on the segmented nodules. Univariate correlation analysis and principal component analysis were performed to select radiomics features. Combined MVD values and selected radiomics features, we conducted a logistic regression classification model. The area under the curve (AUC) was applied to show model performance. Our model reached an AUC value of 0.867 when tested on independent datasets. The performance of the model to differentiate benign, malignant and inflammatory nodules reached AUC values of only 0.908, 0.833, and 0.730, respectively. We conducted a prediction model that shows promising results in distinguishing three different types of lung nodules.

Keywords: Pulmonary nodules, Microvessel density, logistic regression model.

1. Introduction

According to Canada Cancer Society (2021), lung cancer is the primary reason for cancer-related deaths in Canada, accounting for 26% of them. While there has been a decrease in the incidence and mortality rates of lung cancer, the five-year relative survival rate for lung cancer in Canada is still alarmingly low at 19%. This is significantly lower than the overall five-year relative survival rate for all types of cancer combined (67%). These statistics demonstrate the severe impact of lung cancer in Canada and emphasize the need for more efficient prevention strategies, early detection methods, and effective treatments. Further advancements in these areas could increase the survival rate for lung cancer patients and improve health outcomes.

The classification of pulmonary nodules is crucial in determining the appropriate course of action, with malignancy and benignity being the two main categories. Benign nodules can be further divided into non-inflammatory or inflammatory. Accurate classification during early screening is vital for follow-up treatment. Computer tomography (CT) is commonly used to detect even small nodules in the lung and is a standard diagnostic tool for early screening and follow-up observation of lung cancer. However, CT scans have limitations in accurately characterizing lung nodules which often results in invasive procedures like biopsies and surgeries. [1]. These procedures can be expensive and carry significant risks, making it essential to find alternative, less invasive diagnostic and treatment modalities (American Thoracic Society, 2021).

Angiogenesis, the process of forming new blood vessels from pre-existing ones, plays a crucial role in the growth and metastasis of malignant pulmonary nodules and tumors. Microvascular density (MVD), which quantifies the number of microvascular entities present in tumor tissues, is a commonly used approach to assessing the impact of angiogenesis on malignant nodules [2-4]. High MVD values have a significant statistical association with the progression of non-small cell lung cancer (NSCLC) [2, 5]. The high MVD is also associated with metastasis and prognosis. It is important to note that the high MVD results can also be caused by the progression of inflammation,

rather than malignancy, and need to be accurately distinguished before starting any therapeutic intervention [6].

Radiomics is a versatile technique employing computing algorithms to extract otherwise imperceptible quantitative information from baseline CT scans, including basic statistical features and higher-order texture features, facilitating clinical investigations [7]. Radiomics has demonstrated the potential to predict the likelihood of future cancer occurrences and to differentiate malignant pulmonary nodules from benign ones [8].

To date, there are limited studies that have combined MVD with radiomics features to construct a predictive model that differentiates between various types of lung nodules. Thus, the purpose of this project is to develop a radiomics-based machine learning model that can enhance the accuracy and dependability of lung cancer diagnosis, especially when characterizing lung nodules.

2. Materials and Method

2.1. Data Acquisition

This study collected 100 sets of lung CT scans alongside their respective pathological diagnosis results from a hospital between 2010 and 2018. Each patient in the cohort had a single pulmonary nodule confirmed through pathological diagnosis. The group consisted of 58 malignant, 29 inflammatory, and 13 non-inflammatory benign nodules.

A 64-slice multidetector CT scanner performed all CT scanning. The screening parameters were set as follows: tube voltage, 80-140kVp; rotation time, 0.5s; matrix, 512×512; reconstruction section thickness, 1mm and 5mm; scan direction, FFS; effective mAs, reconstruction diameter and table speed were all set automatically.

2.2. Data Processing

Preprocessing operations were applied to standardize the scale and format of the CT scans, thus facilitating their accurate analysis. The first step involved resampling the images to adopt a 1mmx1mmx1mm spatial resolution. Then, laplace enhancement was employed to enhance the texture information, providing visually improved image sharpening. Furthermore, a Gaussian filter was utilized to decrease the noise present within the CT images. These preprocessing techniques ensure the uniformity of the CT images in terms of resolution, format, contrast, and clarity, rendering them appropriate for further analysis and interpretation. Two experienced radiologists performed ROI segmentation, which was then evaluated individually against the pathological diagnosis results to ensure accuracy.

2.3. MVD calculation

Microvessel density (MVD) is typically defined as the number of vessels present in a given unit of analysis volume. To enable a more precise comparison of angiogenesis among the three types of pulmonary nodules, this study calculated three types of MVD according to C. Ushijima et al.'s work [3]. The first “MVD Average” represented the overall MVD and was calculated as the average MVD value of the actual lesion volume. The definition of the second and third types based on the distance between the nodule tissue and normal lung tissue. For nodules with diameters smaller than 20 mm, the central and peripheral volumes’ borders were positioned in the middle of the nodule radius. For nodules larger than 20 mm, the edge was established at the point where the boundary between nodule tissue and normal tissue shrunk 10 mm inwards. “MVD Periphery” pertained to MVD of the actual lesion minus the central portion, with central and peripheral values expressed as percentages that summed to 100%. Automated MVD calculation was performed using a python program.

2.4. Radiomics and Machine Learning Modeling

2.4.1 Radiomics feature extraction

Radiomics features were extracted using PyRadiomics, a python texture extractor implemented with Python [9]. The calculation of radiomics features was based on the following categories with the number of features in parentheses: First order (n=18), gray level co-occurrence matrix (n=24), gray level run length matrix (n=16), gray level size zone matrix (n=16), neighboring gray tone difference matrix (n=5), gray level dependence matrix (n=14), shape (n=14), Laplacian of Gaussian filtered feature (n=186, Sigma=1.0, 3.0, 5.0), first-order wavelet-filtered feature (n=744, Level=1) and 3-D local binary pattern related feature (n=372, Level=2, Radius=1.00, Subdivision=1).

2.4.2 Radiomics feature selection

By utilizing a stratified sampling approach, all 100 patients were assigned randomly to either the training or test set in a 7: 3 ratio. Z-score transformation was employed to standardize the data subsequently. The univariate correlation analysis was then performed using a cutoff threshold of 0.7, after which a primary component analysis (PCA) was conducted with ten components. The primary objective of PCA was to identify the most relevant features.

2.4.3 Machine learning model

MVD values and extracted radiomics features were combined together as the input of machine learning models. The logistic regression model was selected and established for each cohort after comparing with other machine learning models (SVM, decision tree, decision forest, Bayes, KNN).

2.4.4 Statistical Analysis

To assess the discriminatory ability of the model, we obtained the receiver operating characteristic (ROC) curve and the area under curve (AUC). Subsequently, we determined the optimal cutoff value by maximizing the Youden's J index from the ROC curve of the training dataset. This value was then used to calculate the accuracy, sensitivity, and specificity of both the training and test sets. A p-value <0.05 (two-tailed) indicated statistical significance.

3. Results

Illustrated in Figure 1 is an instance of a lung nodule found in one patient. A total of 1409 radiomics features were extracted from their CT dataset, from which ten were selected through various selection procedures. Utilizing both these features and MVD values, we constructed a radiomics model, the efficacy of which was summarized in Table 1. We assessed the diagnostic efficiency of this model on both training and testing datasets using ROC curve analysis (Figure 2). The model exhibited good prognostic ability, with an AUC value of 0.867 in the test dataset. To help us better understand the model's distinguishing capability for each type of nodule, we conducted an analysis of its predictive performance for inflammation, benign, and malignant nodules, and illustrated it in Figure 3. The model had an AUC of 0.908, 0.833, and 0.730 for differentiating benign, malignant, and inflammatory nodules from the other two, respectively. The importance of each selected feature was shown in Figure 4, with three radiomics features and one MVD feature having an importance score that exceeded 80%.

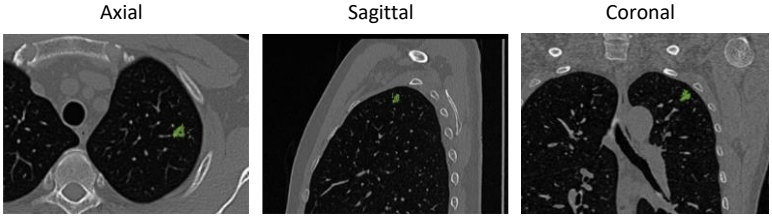

Figure 1. An example of the lung nodule in one patient. From left to right are cross-sectional, coronal, and sagittal views. The specific lung nodule is highlighted in green

Table 1. The overall performance of the logistic regression prediction model

	Train	Test
Accuracy	0.762	0.829
F1_score	0.716	0.778
Recall	0.906	0.875
Precision	0.593	0.700
AUC	0.880 (0.833, 0.921)	0.867 (0.785, 0.937)
Sensitivity	0.906	0.875
Specificity	0.692	0.804

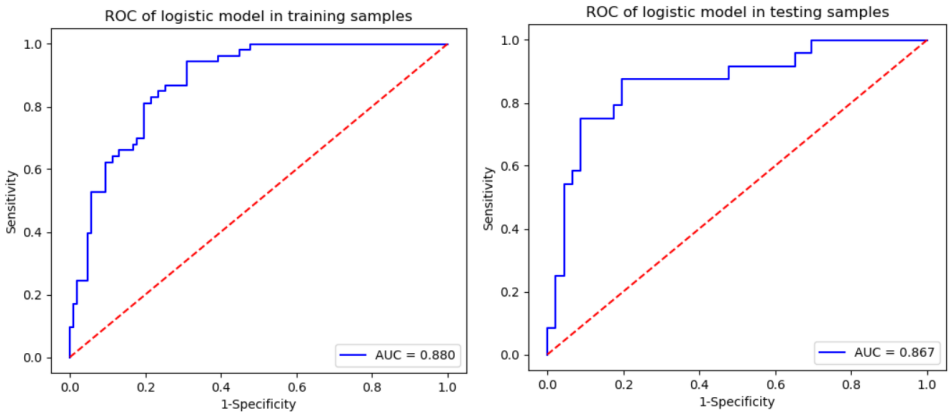

Figure 2. ROC curve of the model in both training and testing datasets

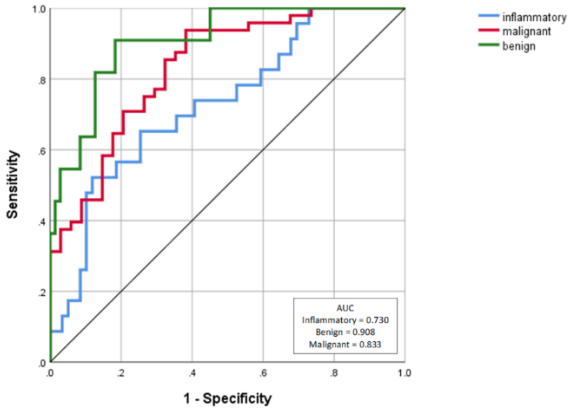

Figure 3. Combined ROC curves for three different types of nodules. These curves are used to assess the predictive ability of the radiomics model for differentiating between these types of nodules

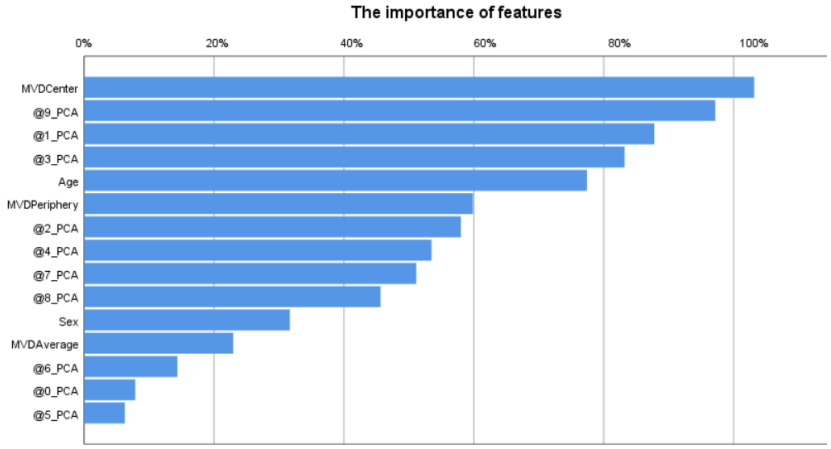

Figure 4. The importance of features for prediction model construction

4. Discussion

Our study focused on the extraction of radiomics features from chest CT sequences, as well as the inclusion of three types of MVD values. By combining these features, we established a machine learning model that can predict the presence of benign, inflammatory, and malignant pulmonary nodules with a high degree of accuracy.

This study incorporated MVD data to further distinguish inflammatory nodules from benign nodules based on the original research. However, the ROC and AUC analyses indicated that the model's ability to differentiate inflammatory nodules was the weakest, with an AUC value of only 0.73. This might be attributed to the fact that the development of inflammation and the growth of malignancies both contribute to increased microvascular growth in nodules, making it more challenging to differentiate them through imaging.

Another important aspect that we discussed in our study is the potential clinical implications of our radiomics model. Physicians can use this model to predict the type of lung nodule non-invasively, allowing for early diagnosis and treatment. This model could also reduce unnecessary invasive procedures and minimize unnecessary exposure to radiation from repeated CT scans. In the longer term, this approach may have the potential to reduce the need for invasive procedures such as biopsies, which can be uncomfortable and carry a risk of complications. It may prevent the progression of malignant nodules and minimize unnecessary medical treatment for benign nodules.

There were several limitations to our study. First, due to the retrospective design of our study, selection bias inevitably existed even with strict inclusion criteria. Second, this was single-center research and the number of patients was relatively small, especially the number of patients with non-inflammatory benign nodules. Finally, more pathological indexes should be included to improve the performance of our model.

In conclusion, we developed a radiomics-based machine learning model that can improve the accuracy and reliability of lung cancer diagnosis, particularly for characterizing lung nodules.

References

[1] Gould MK, Tang T, Liu IL, et al. Recent Trends in the Identification of Incidental Pulmonary Nodules. Am J Respir Crit Care Med. 2015; 192(10): 1208-1214. doi: 10.1164/rccm.201505-0990OC.

[2] Bačić I, Karlo R, Zadro AŠ, Zadro Z, Skitarelić N, Antabak A (2018). Tumor angiogenesis as an important prognostic factor in advanced non-small cell lung cancer (Stage IIIA). Oncol Lett. 2018 Feb; 15(2): 2335-2339. doi: 10.3892/ol.2017.7576. Epub 2017 Dec 8. PMID: 29434942; PMCID: PMC5777107.

[3] Ushijima C, Tsukamoto S, Yamazaki K, Yoshino I, Sugio K, Sugimachi K (2001). High vascularity in the peripheral region of non-small cell lung cancer tissue is associated with tumor progression. Lung Cancer. 2001 Nov; 34(2): 233-41. doi: 10.1016/s0169-5002(01)00246-x. PMID: 11679182.

[4] Yuan A, Yang PC, Yu CJ, et al (1995). Tumor angiogenesis correlates with histologic type and metastasis in non-small-cell lung cancer. Am J Respir Crit Care Med. 1995 Dec; 152(6 Pt 1): 2157-62. doi: 10.1164/ajrccm.152.6.8520790. PMID: 8520790.

[5] Fontanini G, Bigini D, Vignati S, et al (1995). Microvessel count predicts metastatic disease and survival in non-small cell lung cancer. J Pathol. 1995 Sep; 177(1): 57-63. doi: 10.1002/path.1711770110. PMID: 7472781.

[6] Ma SH, Le HB, Jia BH, et al (2008). Peripheral pulmonary nodules: relationship between multi-slice spiral CT perfusion imaging and tumor angiogenesis and VEGF expression. BMC Cancer. 2008 Jun 30; 8: 186. doi: 10.1186/1471-2407-8-186. PMID: 18590539.

[7] Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017; 14(12): 749-62.

[8] Wilson R, Devaraj A. Radiomics of pulmonary nodules and lung cancer. Transl Lung Cancer Res. 2017; 6(1): 86-91.

[9] van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017; 77(21): e104-e107. doi: 10.1158/0008-5472.CAN-17-0339.