**The Use of Radiomics and Machine Learning for Lung Nodule Classification**

Andrei Teleron

Capstone Project Writing and Presentation Final Paper

MScBMI, University of Chicago

**Abstract**

The aim of the project was to build a machine learning algorithm capable of classifying lung nodules as benign or malignant. The LUNGx 2015 Challenge provided 83 lung nodules for use in the model. The CT images were cropped into 64x64 images and non-nodule tissues were removed using image editing software and threshold filters. Nodule size, local binary pattern (LBP), and gray level co-occurrence matrix (GLCM) features were extracted from the images and their values were standardized. A support-vector machine (SVM) algorithm was used for classification. Stratified k-fold cross validation (folds=8) resulted in a mean fold AUC of 0.73. The mean of mean AUCs (following 1000 iterations) stabilized at 0.72.

**Introduction**

Lung cancer has been reported to be the second most common cancer and is the leading cause of cancer death for men and women [1]. The survival rate of patients with metastatic lung cancer is 4% over a period of five years, but 55% for cases in which the cancer is still localized. This, together with the fact that only 16% of lung cancer cases are diagnosed in their early stages [2], emphasizes the importance of early detection and accurate classification between metastatic and non-metastatic types.

Computed tomography (CT) image analysis is the gold standard among radiologists in diagnosing lung cancer, which may manifest as lung nodules. These are anomalous globular tissue that can be benign or malignant depending on its characteristics. The issue occurs when lung cancer is in its early stages, which can make it difficult to detect very small tumors. Radiologists are often mentally burdened and fatigued due to examining many images over the course of a day which may impact his or her ability to correctly find and classify a tumor.

The higher computational abilities of modern computers gave rise to radiomics, a relatively new field that has the power to change modern diagnostic procedures by introducing computer-aided systems that would assist radiologists in diagnosing diseases. This computerized, potentially automated approach with machine learning, is amenable to analyzing large amounts of data—effectively easing the burden of radiologists while giving them more time to communicate with their patients. The system can be more sensitive to tumor heterogeneity while working faster than the human eye [3], and when combined with machine learning, can carry great potential in aiding the work of radiologists.

The 2015 SPIE Medical Imaging Conference (SPIE), in collaboration with the American Association of Physicists in Medicine (AAPM) and the National Cancer Institute (NCI) have issued the SPIE-AAPM LUNGx CT Challenge [4] to promote the creation of quantitative image analysis methods for lung nodule classification. The goal of our study was to explore a machine learning approach in diagnosing lung cancer through the creation of a classification system using support-vector machines using LUNGx dataset. The aim was to have an accuracy comparably better than the proposed solutions submitted to the LUNGx Challenge hosted by the SPIE-AAPM-NCI in 2015 [5].

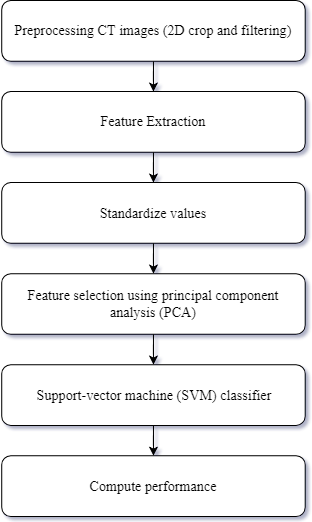
**Methods**

*The Problem*

The informatics problem proposed is, given a dataset containing 70 patients, to build a method or algorithm to determine whether the lung nodules found therein can be classified as benign or malignant. Each patient has at least one lung nodule while others have two for a total of 83 nodules. The data is bundled into calibration, training, and testing sets. Characteristics for the training and testing nodules is illustrated in Table 1. The calibration set containing 10 nodules (5 benign and 5 malignant) was released prior to the training and testing sets to give contestants an idea for the type of data and parameters that would be used. For the purposes of the methodology, the data has been concatenated to increase the sample size—machine learning applications thrive on larger amounts of data. The project pipeline is shown on Figure 1.

**Table 1. Lung nodule characteristics from the LUNGx Challenge (n = 73, calibration data not included)**

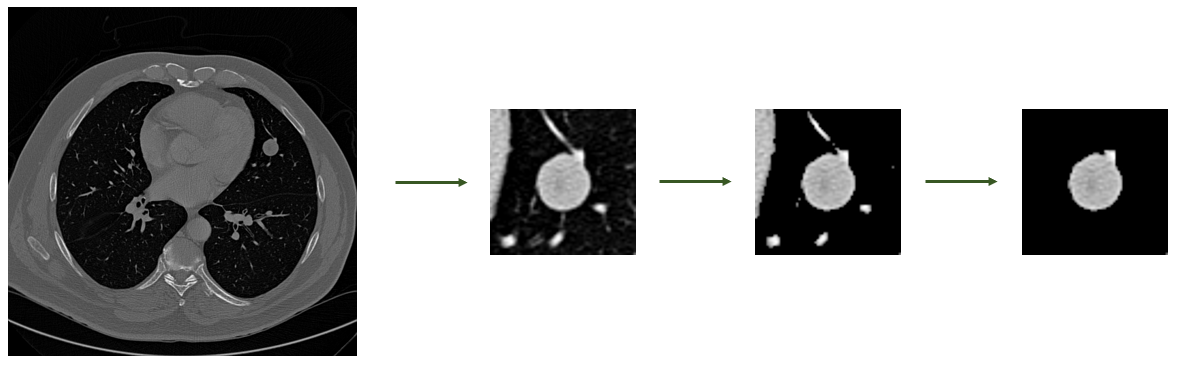
|  |  |  |  |
| --- | --- | --- | --- |
|  | Benign nodules (n = 37) | Malignant nodules (n = 36) | p-value |
| RECIST-based nodule size (mm) |  |  |  |
| Mean (sd) | 18.6 (6.7) | 15.5 (8.2) | 0.12a |
| Median (range) | 13.9 (5.7 - 45.0) | 17.1 (4.6 – 34.6) |  |
| Nodule solidity |  |  |  |
| Nonsolid (n = 4) | 2 (50%) | 2 (50%) |  |
| Part solid (n = 10) | 5 (50%) | 5 (50%) |  |
| Solid (n = 59) | 29 (49%) | 30 (51%) | 0.99b |
| Nodule location |  |  |  |
| Left lower lobe (n = 12) | 5 (42%) | 7 (58%) |  |
| Left upper lobe (n = 23) | 13 (57%) | 10 (43%) |  |
| Right lower lobe (n = 15) | 5 (33%) | 10 (67%) |  |
| Right middle lobe (n = 6) | 2 (33%) | 4 (67%) |  |
| Right upper lobe (n = 17) | 11 (65%) | 6 (35%) | 0.34b |
| Spiculation |  |  |  |
| Absent (n = 12) | 17 (40%) | 26 (60%) |  |
| Present (n = 30) | 19 (63%) | 11 (37%) | 0.045b |
| Note: RECIST = Response Evaluation Criteria in Solid Tumors.  ap-value computed from Student’s t-test.  bp-value computed from chi-square test. | | | |



**Figure 1. Project methodology for lung nodule classification.**

*CT Image Preprocessing*

As with other machine learning methods, the output is only as good as the data that is entered—the 3D CT scans requires preprocessing so that the machine learning classifier can be accurate. An image cropping Python program was created that takes in CT image data (512x512 8-bit axial slices) and the (x, y, instance number) coordinates of each of the lung nodules. When ran, the script was able to automatically and individually crop all 83 lung nodules into 64x64 pixel, grayscale tiff files. Extraneous bodies that weren’t part of the lung nodule were partially removed by running a manual thresholding Python script. GIMP, an open-source image editing software was used to completely remove any remaining non-nodule bodies. Any pixel that did not contain the lung nodule were assigned a value of 0 (black). The preprocessing step is illustrated in Figure 2.



**Figure 2. Preprocessing flow—isolating the lung nodule from other lung structures.**

*Feature Extraction*

After preprocessing the images, features were extracted from the image and placed in a data frame—each row represents a lung nodule and each column represents the extracted features. The first feature was lung nodule size given as the number of pixels in a single image that had a value above 0. Energy (two versions used—LBP and GLCM), contrast, homogeneity, and angular second moment (ASM) are additional features that were extracted. They encode textural data and were derived from the local binary pattern (LBP) histogram of the image as well as the grey level co-occurrence matrix (GLCM) [6]. A total of 6 features were extracted. They were then standardized by removing the mean (centers the values such that the mean for each feature is zero) and scaling to unit variance to reduce bias in the data. For example, features with values in the tens of thousands may be perceived as more significant by the algorithm than features with ranges less than one. Also, having the data appear as standard normally distributed data is a requirement for the radial basis function (RBF) kernel to work correctly in support-vector machine classifiers [7]

*Feature Selection using PCA*

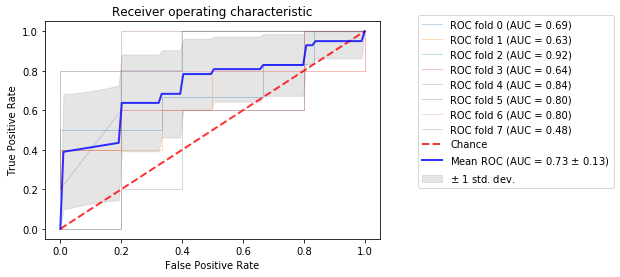
To prevent the issue of overfitting the data and improve the model’s ability to classify new data, the number of features were reduced from 6 to 4 via a process called feature selection [8]. This is done by using principle component analysis (PCA), which projects the data onto linearly uncorrelated vectors called principle components [9]. It is such that the first component vector has the maximum variance while the subsequent component vector also has the maximum possible variance but is constrained in that it is orthogonal to the previous component vector and so on. Once the ranking has been established, the first 4 components were taken, and the rest were discarded. Sklearn’s PCA module, available in Python, was used to perform this operation.

*Machine Learning—Support-Vector Machine Classifier*

After preprocessing the CT image data into a suitable form with a selection of the best features to use, a supervised machine learning method was implemented in the form of SVM to classify pulmonary nodules as benign or malignant. Developed by Vladimir Vapnik in the 1990’s, SVM takes in data which may not be linearly separable at the onset and projects the data into a higher dimensional space where a hyperplane can be drawn to classify the data [10]. The data frame was appended with knowledge of the truth of each nodule (benign – 0, malignant – 1) and was then split into training and testing sets. The method used was stratified k-fold cross validation which splits the data into k-folds where the SVM classifier trains on the k-1 training folds and tests on the remaining fold. The testing fold is moved at the end of each iteration until all the data had a chance to be represented in the testing fold. In each iteration, the SVM algorithm was applied and an accuracy rate was outputted. A range of K values were tested (1-20) to see which value resulted in the highest accuracy rate. The value for K was chosen to be 8 and then the individual values were averaged to obtain a mean accuracy rate for the SVM classifier. The sklearn module in Python was used to implement this methodology.

**Results**

Stratified k-fold cross validation, where the number of folds = 8, results in 10 lung nodules in the testing set and 73 in the training. Figure 3. illustrates the ROC and the AUC for each fold. The mean AUC value for the 8 folds was 73% (13). To demonstrate how stable the model is, 1000 ROCs were generated, and their mean values were averaged together, resulting in 72.3% as the mean.



**Figure 3. Generated ROC curve following stratified K-fold validation. AUC values for each fold and the resulting mean were calculated.**

**Conclusions**

In conclusion, the SVM algorithm was able to beat all the methodologies cited in the paper by Armato III et al. (2016). The use of support-vector machines in lung nodule classification may see potential application in future CAD systems as more accurate systems are developed. The publicly provided data lung nodule data hosted by The Cancer Imaging Archive remains to be a highly regarded tool for researchers in the radiomics field.

**Next Steps**

In Nishio et al.’s (2018) paper, he highlights a similar methodology where his team explores SVM and other machine learning models using 3D images (64x64x64) [11]. It would be beneficial in the future to incorporate 3D images instead of 2D to help increase the algorithm’s performance. The lung image database consortium image collection (LIDC-IDRI) contains 1018 additional lung CT scans that may be used to increase the data count. This endeavor would take a bit longer, since the nodule coordinates are not known, and may need assistance from a trained radiologist. To remedy this, Further exploration into lung nodule segmentation (which could be another project in itself) may also be a next step for the project.

**References**

[1] Key Statistics for Lung Cancer. (2018). Retrieved from <https://www.cancer.org/cancer/non-> small-cell-lung-cancer/about/key-statistics.html

[2] Lung Cancer Fact Sheet. (n.d.). Retrieved from http://www.lung.org/lung-health-and- diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html

[3] Gillies, R. J., PhD, Kinahan, P. E., PhD, & Hricak, H., MD, PhD, Dr(hc). (2015). Radiology, 563-577. doi:10.1148/radiol.2015151169

[4] Cancer Imaging Archive Wiki. (n.d.). Retrieved from https://wiki.cancerimagingarchive.net/display/Public/LUNGx SPIE-AAPM-NCI Lung Nodule Classification Challenge

[5] Armato, S. G., Drukker, K., Li, F., Hadjiiski, L., Tourassi, G. D., Engelmann, R. M., . . . Clarke, L. P. (2016). LUNGx Challenge for computerized lung nodule classification. *Journal of Medical Imaging,* *3*(4), 044506. doi:10.1117/1.jmi.3.4.044506

[6] Sahu, H., & Bhanodia, P. (n.d.). An ANALYSIS OF TEXTURE CLASSIFICATION: LOCAL BINARY PATTERN. *Global Research in Computer Science,* *4*(5). Retrieved from <https://pdfs.semanticscholar.org/494c/7195da3b4e0975085c58122f2b897fe4369c.pdf>.

[7] Sklearn.preprocessing.StandardScaler. (n.d.). Retrieved from <http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>

[8] Wang, S., & Summers, R. M. (2012). Machine Learning and Radiology. *Medical Image Analysis,16*(5), 933-951. doi:https://doi.org/10.1016/j.media.2012.02.005

[9] Jolliffe, I. (2011). Principal Component Analysis. *International Encyclopedia of Statistical Science,*22-32. doi:https://doi.org/10.1007/978-3-642-04898-2\_455

[10] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning,* *20*(3), 273-297. doi:https://doi.org/10.1007/BF00994018

[11] Nishio, M., Nishizawa, M., Sugiyama, O., Kojima, R., Yakami, M., Kuroda, T., & Togashi, K. (2018). Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *Plos One,* *13*(4). doi:10.1371/journal.pone.0195875