Developing a Radiomics Framework for Classifying Non-Small Cell Lung Carcinoma Subtypes

Dongdong Yu[a], Yali Zang[a], Di Dong*[a], Mu Zhou[b], Olivier Gevaert[b], Mengjie Fang[a], Jingyun Shi*[c], and Jie Tian*[a]

[a]The Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[b]The Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University

[c]Department of Radiology, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China

ABSTRACT

Patient-targeted treatment of non-small cell lung carcinoma (NSCLC) has been well documented according to the histologic subtypes over the past decade. In parallel, recent development of quantitative image biomarkers has recently been highlighted as important diagnostic tools to facilitate histological subtype classification. In this study, we present a radiomics analysis that classifies the adenocarcinoma (ADC) and squamous cell carcinoma (SqCC). We extract 52-dimensional, CT-based features (7 statistical features and 45 image texture features) to represent each nodule. We evaluate our approach on a clinical dataset including 324 ADCs and 110 SqCCs patients with CT image scans. Classification of these features is performed with four different machine-learning classifiers including Support Vector Machines with Radial Basis Function kernel (RBF-SVM), Random forest (RF), K-nearest neighbor (KNN), and RUSBoost algorithms. To improve the classifiers' performance, optimal feature subset is selected from the original feature set by using an iterative forward inclusion and backward eliminating algorithm. Extensive experimental results demonstrate that radiomics features achieve encouraging classification results on both complete feature set (AUC=0.89) and optimal feature subset (AUC=0.91).

Keywords: Non-Small Cell Lung Carcinoma, lung nodule, adenocarcinoma, squamous cell carcinoma, feature analysis, computed tomography, classification, computed-aided diagnosis

1. INTRODUCTION

Lung cancer is the leading cause of cancer-related death worldwide, which is classified into two major subtypes, namely, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC is a lethal disease accounting for about 85% of all lung cancers with a dismal 5-year survival rate of 15.9% .[1] NSCLC can be subdivided into adenocarcinoma (ADC), squamous cell carcinoma (SqCC), and other types on the basis of where the lung cancer cell starts from. Over all, ADC and SqCC account for 65% to 70% of all lung cancer patients.

NSCLC patients has been well documented over the past decade to suggest targeted treatment according to the different histologic subtypes.[2,3] The histologic subtypes are discerned by histopathological examination which is a gold standard in the nodule classification. However, the examination approach can be failed, if the tissue sample is inadequate. Computed tomography (CT) has been a major imaging modality for early cancer detection in NSCLC. A majority of image-based studies have been proposed to estimate nodule malignancy likelihood.[4,5] Meanwhile, a promising task is to infer the diagnostic value from CT images, such as the identification of discriminative image features that are able to predict the histologic subtypes. The noninvasive CT image analysis may have complementary roles to histopathologic prediction in lung cancer.

Previous studies have highlighted that the specific proteins and immunohistochemistry for their capability to classify the ADC and SqCC. Ullmann et al. [6] show that protein profiles are feasible tools to classify the lung carcinoma tissue microarrays including 75 ADCs and 67 SqCCs. Non-parametric tests, hierarchical clustering, and principal component analysis are used to analyzed the immunohistochemical expression levels of 86 different

Medical Imaging 2017: Computer-Aided Diagnosis, edited by Samuel G. Armato III, Nicholas A. Petrick,
Proc. of SPIE Vol. 10134, 1013426 · © 2017 SPIE · CCC code: 1605-7422/17/$18 · doi: 10.1117/12.2253923

proteins which were manually scored by pathologists. It shows that the two lung carcinoma subtypes can be predicated with 96% accuracy. However, the immunohistochemistry is time consuming and expensive. In parallel, some studies have shown that the quantitative features from medical images can provide a detailed quantification of tumor characterization. Ha et al.[7] presented an approach to classify ADC and SqCC using a number of texture feature parameters extracted from 18F-fluorodeoxyglucose positron emission tomography scans. More than 200 texture parameters are extracted and fifteen texture features had significant different values between ADC and SqCC. However, the size of patient cohort is 30, which may impede the translational value of the detected image biomarkers.

In this study, we focus on developing computational CT image features for predicting ADC and SqCC. In particular, we introduce a radiomics framework utilizing the image feature analysis for the histopathologic prediction. The presented method includes lung nodule segmentation, imaging feature extraction, feature selection and nodule classification. More specifically, we use the Toboggan Based Growing Automatic Segmentation (TBGA)[8] to segment the lung nodule from the chest CT scans. Then, fifty-two dimensional feature including statistical features and texture features are extracted to characterize the lung nodule. Next, we use four different classifiers: support vector machine with radial basis function kernel (RBF-SVM), random forest (RF), K-nearest neighbor (KNN), RUSBoost classifier to build up the prediction model. Finally, we use the iterative forward inclusion and backward eliminating algorithm to select the significant features and improve the prediction model's ability. The purpose of this study is to investigate the correlation between imaging heterogeneity and histopathologic subtypes of lung cancer and to find the significant imaging features.

2. MATERIALS AND METHODS

2.1 Overview

The flowchart of the proposed framework is shown in Fig. 1. First, the lung nodules treated as volume of interests are segmented from the chest CT image scans using an automatic segmentation method. Second, imaging features are extracted from the lung nodules. Finally, feature selection and supervised classification method with 10-fold cross validation are conducted to build up models for histopathologic prediction.

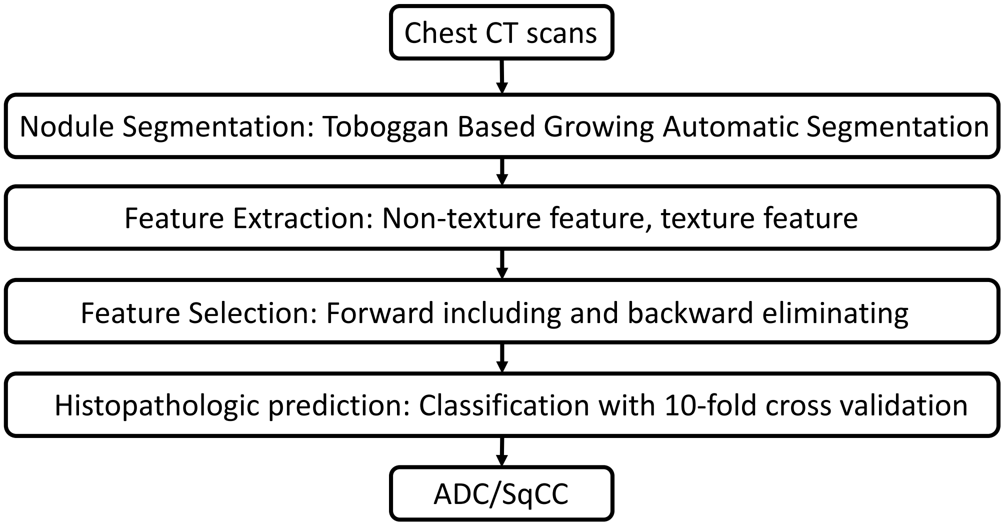

Chest CT scans
Nodule Segmentation: Toboggan Based Growing Automatic Segmentation
Feature Extraction: Non-texture feature, texture feature
Feature Selection: Forward including and backward eliminating
Histopathologic prediction: Classification with 10-fold cross validation
ADC/SqCC

Figure 1. Flowchart of the proposed histopathologic prediction framework.

2.2 Patient cohort and volume of interest

The patient cohort consists of 434 patients including 324 ADC and 110 SqCC patients. Both non-enhanced and contrast-enhanced chest CT images are acquired on Philips Brilliance 40 and Siemens Defintion AS. The

Table 1. Specific texture feature parameters. Abbreviations of feature names: Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray-Level Nonuniformity (GLN),Run-Length Nonuniformity (RLN), Run Percentage (RP), Low Gray-Level Run Emphasis (LGRE), High Gray-Level Run Emphasis (HGRE),Short Run Low Gray-Level Emphasis (SRLGE), Short Run High Gray-Level Emphasis (SRHGE), Long Run Low Gray-Level Emphasis (LRLGE), Long Run High Gray-Level Emphasis (LRHGE), Gray-Level Variance (GLV), Run-Length Variance (RLV), Small Zone Emphasis (SZE), Large Zone Emphasis (LZE), Gray-Level Nonuniformity (GLN), Zone-Size Nonuniformity (ZSN), Zone Percentage (ZP), Gray-Level Variance (GLV), Zone-Size Variance (ZSV), Low Gray-Level Zone Emphasis (LGZE), High Gray-Level Zone Emphasis (HGZE), Small Zone Low Gray-Level Emphasis (SZLGE), Small Zone High Gray-Level Emphasis (SZHGE), Large Zone Low Gray-Level Emphasis (LZLGE), Large Zone High Gray-Level Emphasis (LZHGE).

Texture Type	Texture Subtype	Texture Index
First Order	Global feature (#5)	Variance, Skewness, Kurtosis, Entropy, Uniformity
Second Order	GLCM texture feature (#9)	Contrast, Energy, Variance, Average, Correlation, Homogeneity, Entropy, Dissimilarity, IDM
High Order	GLRLM texture feature (#13)	SRE, LRE, GLN, RLN, RP, LGRE, HGRE, SRLGE, SRHGE, LRLGE, LRHGE, GLV,RLV
	GLSZM texture feature (#13)	SZE, LZE, GLN, ZSN, ZP, GLV, SZV, LGZE, HGZE, SZLGE, SZHGE, LZLGE, LZHGE
	NGTDM texture feature (#5)	Coarseness, Contrast, Busyness, Complexity, Strength

acquisition parameters of Philips Brilliance 40 are as follows: rotation time = 0.75s, detector collimation = 32*1.25mm, field of view (FOV) = 300 * 300mm, pixel matrix = 512 * 512, Filter sharp (C) for CT reconstruction, while the Siemens Defination AS is with the following acquisition parameters: rotation time = 0.5s, detector collimation = 64 * 0.625mm, FOV = 300 * 300mm, image matrix = 512 * 512, kernel B31f medium sharp+ for CT reconstruction. The spacing of x ranges from 0.53 to 0.89mm, the spacing of y ranges from 0.53 to 0.89mm, the spacing of z is a fixed value 0.7mm. To eliminate the effect of image resolution, all the nodule images from both datasets are resampled and set the resolution to a fixed 0.8 mm per pixel along all three axes.

In this study, a robust and automatic 3D segmentation method named Toboggan Based Growing Automatic Segmentation (TBGA) [8] is used to segment the lung nodules.

2.3 Feature extraction

In order to characterize the nodules, we extract fifty-two 3D image features. The features used to characterize the delineated tumor are described below.

2.3.1 Non-texture analysis

We extract the following seven feature parameters from the segmented lung nodules of each patients CT scans: IntensityMax, IntensityMin, IntensityAve, IntensityStd, Volume, Solidity, and Eccentricity. IntensityMax, IntensityMin, IntensityAve and IntensityStd are the maximum, minimum, average, and standard deviation intensity value of the nodule, respectively. Volume stands for the nodule size. Solidity indicates the ratio of the number of voxels in the nodule to the number of voxels in the 3D convex hull of the nodule. Eccentricity denotes the ellipsoid which best fit the nodule.

2.3.2 Texture analysis

We extract forty-five texture feature parameters including the first order statistics, second order statistics and higher order statistics. Five first order statistics parameters including: Variance, Skewness, Kurtosis, Entropy, Uniformity are extracted to describe the intensity histogram distribution of the nodule region. Nine second order statistics parameters can be calculated from the Gray Level Co-occurrence Matrix (GLCM) .[9] Other thirty-one high order feature parameters are calculated from the Gray Level Size Zone Matrix (GLSZM),[10] Gray Level Run Length Matrix (GLRLM) ,[11] and Neighborhood Gray Tone Difference Matrix (NGTDM) .[12] All of the GLCM, GLSZM, GLRLM, and NGTDM based texture feature parameters are calculated using 3D analysis. The specific texture features are listed in the Tab. 1.

2.4 Feature Selection

In this study, we use the iterative forward including and backward elimination to find the optimal feature set to characterize the difference between ADC and SqCC. Starting from an empty feature set, iterative forward inclusion and backward elimination [13] are employed to include and eliminate feature attribute in the current feature set to increase the cost function. The cost function ACC is defined as below:

$$cost = \frac{1}{2} * (\frac{TP}{TP+FN} + \frac{TN}{FP+TN}), \tag{1}$$

TP stands that adenocarcinoma patients correctly identified as adenocarcinoma, FP denotes that squamous cell carcinoma patients incorrectly identified as adenocarcinoma, TN refers that squamous cell carcinoma patients incorrectly identified as squamous cell carcinoma, and FN indicates that adenocarcinoma patients incorrectly identified as squamous cell carcinoma.

2.5 Histopathologic prediction

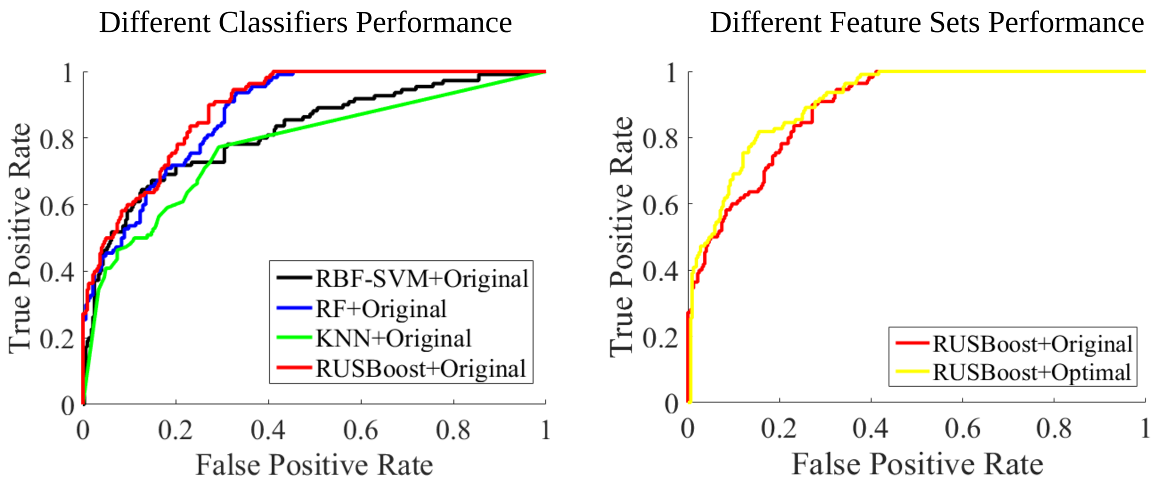
The original image feature set which is used for histologic subtype classification consisted of 432 images (324 ADC and 110 SqCC) is described in Sec. 2.3. For the classification model development, we adopt four different algorithms: support vector machine with radial basis function kernel (RBF-SVM) ,[14] random forest (RF) ,[15] K-nearest neighbor (KNN) ,[16] and RUSBoost algorithms .[17] In additional, we also build up the classification model using the optimal feature set selected by the iterative forward inclusion and backward elimination. The original image feature set and the optimal feature set are both trained with the four different classifiers with the 10-fold cross validation. We report the area under the receiver operating characteristic (ROC) curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and geometric mean (GM) as metrics to assess the performance of the supervised classification from 100 times of 10-fold cross validation. All the metrics are defined as below, and TP, TN, FP, and FN are described in Sec. 2.4,

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+FN+FP+TN} \\ Sensitivity &= \frac{TP}{TP+FN} \\ Specificity &= \frac{TN}{FP+TN} \\ PPV &= \frac{TP}{TP+FP} \\ NPV &= \frac{TN}{TN+FN} \\ GM &= \sqrt{Sensitivity * Specificity}. \end{aligned} \tag{2}$$

3. RESULTS AND DISCUSSION

To build up a robust histologic classification model, we report average results from 100 times 10-fold cross validation on the patient cohort (324 ADCs and 110 SqCCs) with four different classifiers and two different feature sets. We use the AUC and GM as the main rules to measure the performance of classification. AUC is a generalized indicator of the classifier that is independent of the sample class distribution, and the GM maximizes the accuracy on each of the two classes while keeping these accuracies balanced.[18]

Tab. 2 shows the classification results with different classifiers with regards to two feature sets. In this study, we find that the classification model by using the RUSBoost classifier best predict the adenocarcinoma (ADC) and squamous cell carcinoma (SqCC). RUSBoost is the best performing classifier which handles the dataset best among all the four different classifiers. Also, Tab. 2 shows the classification with the original feature set and the optimal feature set selected from the original feature set using iterative forward inclusion and backward elimination by using RUSBoost classifier. By using the optimal feature set with selected 20 features, the performance achieves an average accuracy of 81.5%, sensitivity of 82.6%, specificity of 78.3%, geometric



(a) (b)

Figure 2. ROC curves of lung nodule histologic prediction using different types of classification algorithms based on original feature set (a): the black line, blue line, green line, and red line respectively indicate the ROC prediction curve by using RBF-SVM, RF, KNN, and RUSBoost. ROC curves of lung nodule histologic prediction using original feature set and optimal feature set based on RUSBoost classifier (b): the red line and the yellow line indicate the ROC prediction curves by using original feature set and the optimal feature set.

Table 2. Comparison of different classifiers for histologic prediction using different feature sets.

Classifiers	Feature Set	Accuracy	Sensitivity	Specificity	PPV	NPV	GM	AUC
RBF-SVM	Original	79.8%	**97.6%**	27.6%	79.9%	**79.5%**	0.52	0.82
RF	Original	81.0%	91.5%	50.1%	84.4%	66.7%	0.68	0.88
KNN	Original	77.2%	86.1%	51.1%	83.8%	55.5%	0.66	0.78
RUSBoost	Original	78.8%	80.2%	74.6%	90.3%	56.1%	0.77	0.89
RUSBoost	Optimal	**81.5%**	82.6%	**78.3%**	**91.8%**	60.5%	**0.80**	**0.91**

mean of 0.80, and AUC of 0.91. With the RUSboost classification model, the GM and AUC of the model using optimal feature set outperform the original feature set by approximated 3.9% and 2.2%. We can see that the classification model built up by the optimal feature set outperformed the original feature set. This indicates that feature selection can be efficient to eliminate potential irrelevant features and improve prediction performance. Feature selection leads to a selected optimal feature set with 20-dimensional features. More specifically, the optimal feature set includes a statistical feature of IntensityMin and 19-dimensional texture features (3 first-order texture features, 2 second-order texture features, and 14 high-order texture features). We also report ROC curves in Fig. 2 to fully observe the classification outcomes.

4. CONCLUSION

In this paper, we investigate the association between CT imaging features and histologic subtypes for patients suffering from NSCLC. The proposed radiomics analytic framework presents encouraging results in predicting the adenocarcinoma and squamous cell carcinoma by extracting 52-dimensional radiomics feature. In particular, we achieve the highest classification results with AUC of 0.91 by applying the RUSBoost classifier with 20 selected radiomics features. This study based on radiomics analysis and histopathlogical characteristics supports the potential of computational CT-based analysis as an non-invasive means to facilitate NSCLC diagnosis. The proposed prediction model therefore holds promise to provide objective and reproducible diagnosis for non-small cell lung carcinoma.

5. ACKNOWLEDGMENT

This work is supported by the National Basic Research Program of China under Grant 2011CB707700, the National Natural Science Foundation of China under Grant Nos. 81227901, 61231004, 61363049, the Chinese Academy of Sciences Fellowship for Young International Scientists under Grant 2013Y1GB0005, the National High Technology Research and Development Program of China (863 Program) under 2012AA021105, the Guangdong Province-Chinese Academy of Sciences comprehensive strategic cooperation program under 2010A090100032 and 2012B090400039, the Beijing Natural Science Foundation under Grant No. 4132080, the NSFC-NIH Biomedical collaborative research program under 81261120414, the Fundamental Research Funds for the Central Universities under Grant No. 2013JBZ014, the National Basic Research Program of China under Grant No. 61301002 and the Open Project Program of the State Key Laboratory of Management and Control for Complex Systems under Grant No. 20140101. O.G. is supported by the National Institute Of Biomedical Imaging And Bioengineering of the National Institutes of Health under Award Number R01EB020527. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

[1] Kligerman, S. and White, C., "Epidemiology of lung cancer in women: risk factors, survival, and screening," *American Journal of Roentgenology* **196**(2), 287–295 (2011).

[2] Dubey, S. and Powell, C. A., "Update in lung cancer 2008," *American journal of respiratory and critical care medicine* **179**(10), 860–868 (2009).

[3] Sandler, A., Gray, R., Perry, M. C., Brahmer, J., Schiller, J. H., Dowlati, A., Lilenbaum, R., and Johnson, D. H., "Paclitaxel–carboplatin alone or with bevacizumab for non–small-cell lung cancer," *New England Journal of Medicine* **355**(24), 2542–2550 (2006).

[4] El-Baz, A., Nitzken, M., Khalifa, F., Elnakib, A., Gimelfarb, G., Falk, R., and El-Ghar, M. A., "3d shape analysis for early diagnosis of malignant lung nodules," in [*Information Processing in Medical Imaging*], 772–783, Springer (2011).

[5] Shen, W., Zhou, M., Yang, F., Yang, C., and Tian, J., "Multi-scale convolutional neural networks for lung nodule classification," in [*Information Processing in Medical Imaging*], 588–599, Springer (2015).

[6] Ullmann, R., Morbini, P., Halbwedl, I., Bongiovanni, M., Gogg-Kammerer, M., Papotti, M., Gabor, S., Renner, H., and Popper, H. H., "Protein expression profiles in adenocarcinomas and squamous cell carcinomas of the lung generated using tissue microarrays," *The Journal of pathology* **203**(3), 798–807 (2004).

[7] Ha, S., Choi, H., Cheon, G. J., Kang, K. W., Chung, J.-K., Kim, E. E., and Lee, D. S., "Autoclustering of non-small cell lung carcinoma subtypes on 18f-fdg pet using texture analysis: a preliminary result," *Nuclear medicine and molecular imaging* **48**(4), 278–286 (2014).

[8] Song, J., Yang, C., Fan, L., Wang, K., Yang, F., Liu, S., and Tian, J., "Lung lesion extraction using a toboggan based growing automatic segmentation approach," (2015).

[9] Haralick, R. M., Shanmugam, K., and Dinstein, I. H., "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on* (6), 610–621 (1973).

[10] Thibault, G., Fertil, B., Navarro, C., Pereira, S., Cau, P., Levy, N., Sequeira, J., and Mari, J., "Texture indexes and gray level size zone matrix application to cell nuclei classification," (2009).

[11] Dasarathy, B. V. and Holder, E. B., "Image characterizations based on joint gray levelrun length distributions," *Pattern Recognition Letters* **12**(8), 497–502 (1991).

[12] Amadasun, M. and King, R., "Textural features corresponding to textural properties," *Systems, Man and Cybernetics, IEEE Transactions on* **19**(5), 1264–1274 (1989).

[13] Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M., "An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations," *Nature genetics* **44**(7), 825–830 (2012).

[14] YAN, X.-f., GE, H.-w., and YAN, Q.-s., "Svm with rbf kernel and its application research," *Computer Engineering and Design* **11**(27), 1996–1998 (2006).

[15] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P., “Random forest: a classification and regression tool for compound classification and qsar modeling,” *Journal of chemical information and computer sciences* **43**(6), 1947–1958 (2003).

[16] Cunningham, P. and Delany, S. J., “k-nearest neighbour classifiers,” *Multiple Classifier Systems* , 1–17 (2007).

[17] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., and Napolitano, A., “Rusboost: A hybrid approach to alleviating class imbalance,” *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* **40**(1), 185–197 (2010).

[18] Barandela, R., Sánchez, J. S., Garcıa, V., and Rangel, E., “Strategies for learning in class imbalance problems,” *Pattern Recognition* **36**(3), 849–851 (2003).