Automatic Annotation of Radiological Observations in Liver CT Images

Francisco Gimenez[1], Jiajing Xu[2], Yi Liu[1], Tiffany Liu[1], Christopher Beaulieu[3], Daniel Rubin[3*], Sandy Napel[3*]

[1]Biomedical Informatics Training Program, Stanford University, CA; [2]Department of Electrical Engineering, Stanford University, CA; [3]Department of Radiology, Stanford University, CA

Abstract

We aim to predict radiological observations using computationally-derived imaging features extracted from computed tomography (CT) images. We created a dataset of 79 CT images containing liver lesions identified and annotated by a radiologist using a controlled vocabulary of 76 semantic terms. Computationally-derived features were extracted describing intensity, texture, shape, and edge sharpness. Traditional logistic regression was compared to L_1-regularized logistic regression (LASSO) in order to predict the radiological observations using computational features. The approach was evaluated by leave one out cross-validation. Informative radiological observations such as lesion enhancement, hypervascular attenuation, and homogeneous retention were predicted well by computational features. By exploiting relationships between computational and semantic features, this approach could lead to more accurate and efficient radiology reporting.

Introduction

Liver lesions stem from a variety of causes and manifest in several variations on CT images; some are benign while others may be malignant, and the different diagnoses demonstrate a variety of visual appearances. The ability to differentiate these lesions efficiently and accurately is important to patient treatment and outcome. Contrast-enhanced CT imaging is the dominant technology used for liver lesion diagnosis [1]. This modality takes advantage of the fact that the liver receives blood from two main sources, the portal vein and the hepatic artery. The portal vein supplies about 80% of blood to the liver with the hepatic artery providing the other 20%. Due to varying physiology among liver lesions, differing lesion types may not share the same blood intake proportions as the surrounding liver tissue. Multi-phasic contrast-enhanced imaging takes advantage of this by obtaining images of the liver at multiple time points after injection of contrast agent. This allows for visualization of lesions due to the difference in time between the arrival of contrast agent in the hepatic and portal circulations. This difference causes several distinctive imaging features on contrast-enhanced imaging. As an example, primary liver cancer tumors receive all blood from the hepatic artery, so they contain higher concentrations of contrast agent than surrounding liver parenchyma during the arterial phase [2,3]. Arterial-phase contrast-enhanced CT, therefore, may be helpful in finding masses that exhibit malignant tumor characteristics. Other phases may be useful for differentiating tumor types. For example, metastatic tumors appear less dense compared to the normal liver during the portal venous phase.

The difference in density between a lesion and its surrounding tissue at various times after the injection of iodine in a peripheral vein in multi-phasic imaging is called the temporal enhancement pattern. Analysis of a lesion's temporal enhancement pattern through the different phases of image acquisition helps radiologists to make diagnoses. Unfortunately, the specificity of this method is a function of the size of the lesion and prone to a high false positive rate because several types of liver lesions, including benign ones, have similar manifestations on CT images [4]. Human variability has also been shown to be a challenge for lesion differential diagnosis, and automated methods are being investigated to improve diagnosis of these lesions [5].

Recent research has investigated computer-aided methods to support diagnosis by providing a database of annotated images that can be retrieved by similarity[6] and further indicates that radiological observations drawn from a controlled vocabulary can lead to accurate diagnoses of liver tumor types from CT images[7]. These observations can be treated as features of the image derived from human semantic annotations. As a result, we refer to them as *semantic features.* These semantic features allow radiologists to make their observations consistent, explicit, and machine-accessible. Radiological tools such as the ePAD (formerly iPad) [8] have been implemented with the purpose of recording these annotations in a facile manner.

[*] These authors contributed equally to this work

Computational analysis of these images may overcome this issue by creating quantitative and unbiased descriptors of image features. These computational features can be computed directly from the image's pixel values, independent of the semantic features. Digital image processing techniques have been used to extract features indicative of lesion attenuation, texture [9,10], edge and shape [11-14]. We hypothesize that these computational features can be coupled with methods in statistical machine learning to form a framework to predict semantic features. This could be useful in building a decision support system to aid radiology interpretation.

In this study, we compare the performance of two widely used machine-learning methods to achieve this: logistic regression and L_1-regularized logistic regression (LASSO). From these results, we can determine which semantic features are predicted by computational ones, and which computational features are the most useful for this purpose. Knowledge of which semantic features are not predicted well could lead to the development of new computational features for this purpose. Ultimately, prediction of semantic features from computational ones could lead to reduced variability in image interpretation.

Materials and Methods

Dataset

With IRB approval, we obtained 79 de-identified CT images of liver lesions in the portal venous phase, including eight types of lesion diagnoses: metastasis, hemangioma (abnormal buildup of blood vessels in the liver), hepatocellular carcinoma, focal nodular hyperplasia (unknown cause), abscess (inflammation), laceration (injury or tear), fat deposition and cyst. For the initial development of image processing algorithms, we chose to focus on the most commonly used phase of imaging, the portal venous phase. Later studies will incorporate unenhanced, arterial, and delayed-phase images, which are important in the diagnosis of many liver lesions. For each scan, the axial slice with the largest lesion area was selected for analysis. A radiologist drew and recorded a Region of Interest (ROI) around the lesion on these images using the freely available OsiriX workstation [15].

Semantic Features

A radiologist annotated the ROI with the Electronic Physician's Annotation Device (ePAD; formerly iPAD) [8]. The ePAD system is based on a controlled radiological vocabulary called RadLex to define semantic features [16], and enforces complete description of the required aspects of the visualized lesion. We extended the RadLex terminology to include a broader array of descriptive terms for this study that more comprehensively describe liver lesions. Each annotation resulted in the creation of a binary semantic feature vector of length 76 to indicate positive or negative observations.

Semantic features were not all equally likely and ranged widely in annotation frequency. The entropy of each semantic feature was measured using the binary entropy function in order to determine the distribution of positive versus negative observations. The binary entropy function is defined as:

$$H(X) = -p\log_2 p - (1-p)\log_2(1-p)$$

Where:

Table 1: Semantic features with entropy above 0.4

Semantic Feature	Binary Entropy
smooth margin	0.9971
ovoid	0.9971
normal perilesional tissue	0.9804
homogeneous	0.9804
heterogeneous	0.9738
enhancing	0.9738
solitary lesion	0.9484
nonenhancing	0.914
homogeneous enhancement	0.8859
hypodense	0.8702
circumscribed margin	0.8702
round	0.8163
multiple lesions 1-5	0.8163
lobular	0.8163
homogeneous fade	0.727
multiple lesions 6-10	0.7012
peripheral discontinuous nodular enhancement	0.6739
multiple lesions >10	0.6739
homogeneous retention	0.6739
centripetal fill-in	0.6739
soft tissue density	0.6451
poorly-defined margin	0.6451
abuts capsule of liver	0.6451
water density	0.5822
hypervascular	0.5822
irregularly shaped	0.548
irregular margin	0.548
heterogeneous enhancement	0.548
lobulated margin	0.5116
internal nodules	0.5116

$$p = \frac{\#\ \text{of Positive Observations}}{\text{Total}\ \#\ \text{of Observations}}$$

H(X) ranges from 0 to 1, achieving a maximum when a semantic feature has an equal number of positive and negative observations [17]. To avoid degenerate classification cases, only features with entropy greater than 0.4 were considered for classification, resulting in a total 30 semantic features per lesion (Table 1).

Computational Features

The pixel data within the segmented liver lesions were processed to quantify contrast, texture, boundary, and shape (Table 2). Each lesion's extracted computational features were concatenated into a 431-dimensional feature vector.

Table 2: Computational features and their dimensions

Computational Feature Group	Dimension
Contrast	**2**
Proportion of pixels with intensity larger than 1100	1
Difference of means	1
Texture	**349**
Histogram	9
Histogram - Peak Position	1
Histogram - Entropy	1
Histogram - Haar	1
Histogram - Daubechies	324
Variance	1
Gabor	12
Edge	**61**
Edge Sharpness	60
Histogram on Edge	1
Shape	**19**
Compactness	1
Roughness	1
Local Area Integral Invariant	15
Radial Distance Signature	2
All Features	**431**

Contrast Features: 2-element feature vector containing: (a) the proportion of pixels with intensity larger than 1100 Hounsfield Units (HU) and (b) the difference in the mean intensity values for pixels inside the lesion and within a 5-pixel rim outside the liver lesion.

Texture Features: 349-element feature vector containing: (a) 13-element gray-level histogram-based, including the 9-bin histogram itself, the low frequency coefficients of its 3-level Haar wavelet transform, the abscissa of its peak, entropy, and its variance [9], (b) 12-element Gabor features [10] including the mean of the Gabor energy in the frequency domain over 3 scales and 4 orientations in a total of 12 bins, and (c) 324-element Daubechies features with the dominant sub-band in a 2-scale Daubechies wavelet transform [18].

Margin Sharpness Features: 61-element feature vector computed as follows: (a) We recorded the image intensity values along normals to the lesion contour at multiple points and then fit a sigmoid function to these values. Two parameters for the fitted sigmoid, scale and window, were used to characterize each line segment. The scale measures the difference in intensities outside and inside the lesion, and the window measures the width of the transition from the liver lesion to the surrounding normal liver at the boundary. Two 30-bin histograms for the scale and window parameters were then created to form a 60-element feature vector. (b) We also recorded the number of modes in the histogram of all pixels recorded from each normal [14].

Shape Features: 19-element feature vector describing: (a) compactness [19], (b) roughness [20], (c) local area integral invariant descriptor including the mean and standard deviation for 5 different scales [11,12], (d) radial distance signatures including mean and standard deviation [13].

Classification

Logistic regression was used to calculate the probability of a computational feature vector representing a positive or negative semantic feature. In order to mitigate the potential for over-fitting due to a large number of predictors (431 computational features) compared to data points (79 lesions), we also used LASSO [21] to weight features given sparsity in the computational feature set. We measured classifier accuracy with leave-one-out cross-validation (LOOCV).

Receiver operating characteristic (ROC) curves were calculated over the probability of a semantic feature being recorded. The resulting areas under the curves (AUC) were measured to quantify the predictive value of these classifiers. Classification accuracy for a probability threshold was measured by:

$$Accuracy = \frac{\text{True Positives } + \text{ True Negatives}}{\text{Total number of lesions}}$$

Thresholds were determined empirically over a range from 0 to 1. These thresholds were then used to calculate the misclassification rate of a classifier for a semantic feature.

All programming was performed in MATLAB. LASSO classification was done using the glmnet package for MATLAB [22].

Results

Classifier Performance

Figure 1 (a) shows the area under the ROC curve (AUC) and (b) shows the misclassification rate (MCR) of LASSO and logistic regression. Table 3 shows the mean and standard deviations of the AUC and MCR.

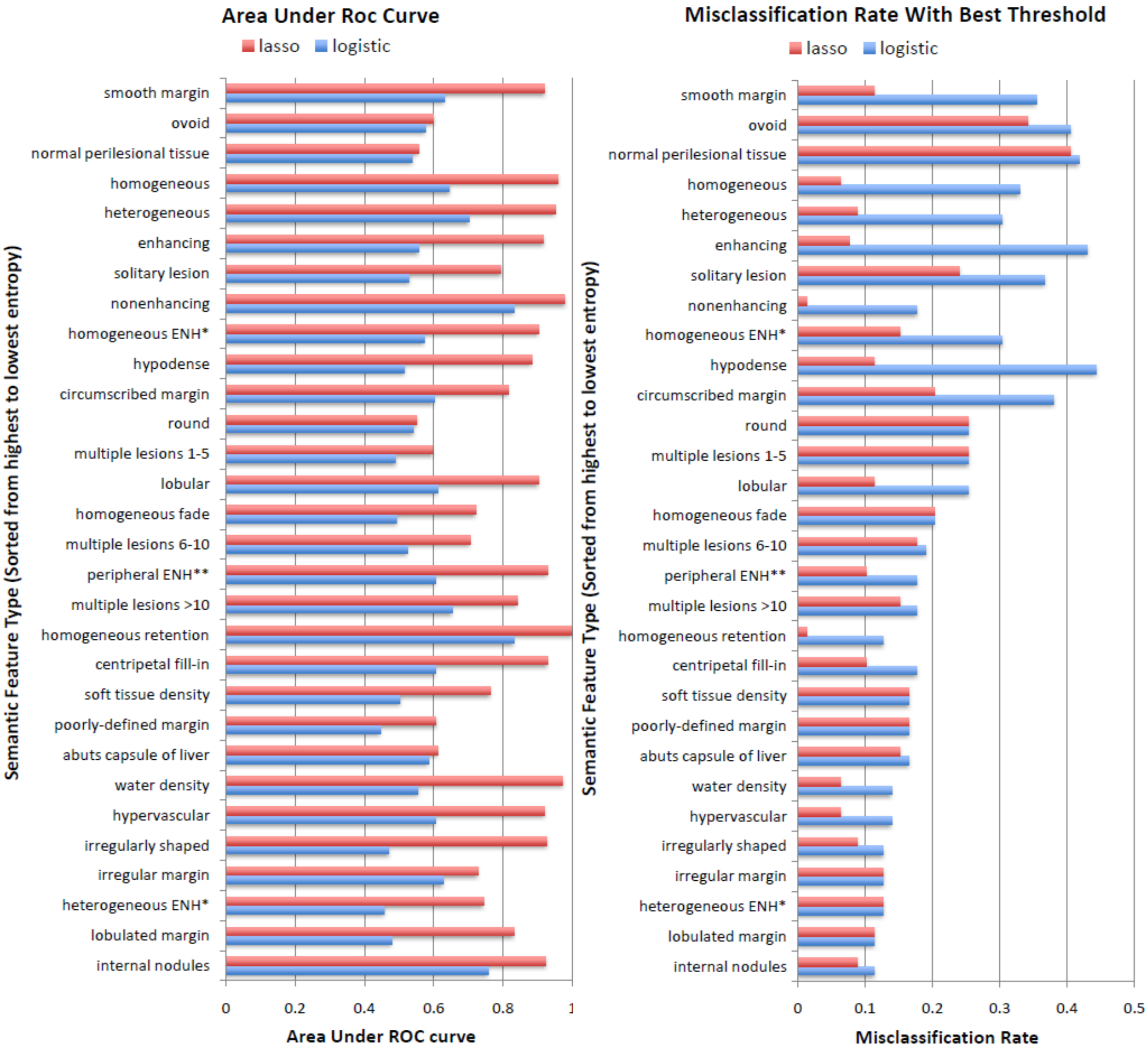

Figure 1: (a) Calculated area under the ROC curve given the probability of a semantic feature occuring for each lesion. (b) Misclassification rate using optimally determined threshold. Semantic features are ordered from highest to lowest entropy. ENH* = enhancement, ENH** = discontinuous nodular enhancement.

Table 3: Aggregate classification statistics

Area under curve statistics

Classifier	Mean	Standard Deviation
LASSO	0.816	0.141
Logistic	0.584	0.098

Misclassification rate statistics

Classifier	Mean	Standard Deviation
LASSO	0.1443	0.0881
Logistic	0.2367	0.1085

We used a two-sample t-test to compare the AUC and MCR of the classifiers. The difference in classifier performance was statistically significant using both AUC and MCR as evaluation metrics ($p_{auc} = 7.2*10^{-10}$ and $p_{mcr} = 6.2*10^{-4}$). Both metrics indicated that LASSO was a stronger classifier.

Semantic Features of Interest

Using the LASSO classifier, several semantic features were shown to be well predicted given computation features. Five were found to have an AUC greater than 0.95: water density, homogeneous retention, non-enhancing, heterogenous, and homogenous.

Conversely, four semantic features were shown to have an AUC under 0.6: multiple lesions 1-5, round, normal perilesional tissue, and ovoid.

There were interesting results with regards to which semantic features were predictable. While we had developed several computational features to quantify shape, two of the most difficult semantic features to predict were round and ovoid. One possible reason for this discrepancy between computational and semantic features is that round and ovoid are inherently subjective terms. Hence, there might exist human variability in such descriptors that cannot be accounted for computationally. This explanation lends credence to further use of computational features for characterizing lesion shape, as there is no variability in our methods.

Another interesting result is our system's ability to predict seemingly impossible semantic features from one lesion. For example, "multiple lesions > 10" was very well-predicted even though analysis was carried out on only one lesion. Such results might be possible because there is an explanatory disease behind both lesion morphology and multiple lesions. Thus, we can indirectly predict number of lesions based on analysis of a single lesion's physical characteristics alone.

Computational Feature Analysis

Computational features were fitted to each semantic feature vector using the LASSO model. Each fit produced a 431-dimensional set of weights for the computational features. Features with large magnitude weights were deemed most informative. The L_1 norm regularization in the model imposes a sparse weight selection; most features have zero weight. To quantify the model complexities, we fit a lasso model to each semantic feature group and counted the number of non-zero coefficients. This corresponds to the number of relevant computational features in each semantic feature group. On average, each semantic feature only employed 12.6 (± 4.3) computational features. Moreover, of the entire set of 431 computational features, only 126 computational features had non-zero for *any* semantic feature vector.

Computational features that consistently had high magnitude weights were considered as characteristic features of these lesions. Characteristic features were found by taking the sum of the absolute value of weights across all semantic features. The 20 features with highest sum of weight magnitudes were categorized according to their associated algorithms. Daubechies Wavelets, Edge Sharpness, Gabor Transform, and the Local Area Invariant Descriptor were found to be the most informative feature groups.

Discussion

In this study we present a framework for predicting radiological observations of liver lesions using computational image features. We computed a wide array of computational features from CT images of liver lesions and used these features to train logistic regression and LASSO classifiers. We experimented with other classifiers such as k-nearestneighbor, support vector machines, and latent discriminant analysis. These methods were severely hampered by such high dimension/low sample sized data, either not reaching any model convergence in training phase, or demonstrating poor results. As a result, we only focused on unregularized and regularized logistic regression. We evaluated the classification results using the area under the ROC curve and misclassification accuracy as metrics. From these results we were able to establish correlations between our computational feature set and radiological observations.

Our approach can be used to evaluate the predictive value of computational features as well as to determine radiological observations that are difficult to predict from computational image features. While computational analysis is not likely to replace the trained eyes of a radiologist, this work can be used to develop decision support tools to increase accuracy and efficiency of radiological diagnosis.

References

1. Baron RL. Understanding and optimizing use of contrast material for CT of the liver. AJR Am J Roentgenol. 1994 Aug. 1;163(2):323–331.

2. Lautt WW, Greenway CV. Conceptual review of the hepatic vascular bed. Hepatology. W.B. Saunders; 1987;7(5):952–963.

3. Matsui O, Kadoya M, Kameyama T, Yoshikawa J, Takashima T, Nakanuma Y, et al. Benign and malignant nodules in cirrhotic livers: distinction based on blood supply. Radiology. 1991 Jan. 1;178(2):493–497.

4. Lencioni R, Cioni D, Pina della C, Crocetti L, Bartolozzi C. Imaging diagnosis. Semin Liver Dis. 2005;25(2):162–170.

5. Armato SG, McNitt-Gray MF, Reeves AP, Meyer CR, McLennan G, Aberle DR, et al. The Lung Image Database Consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans. Acad Radiol. 2007 Nov. 1;14(11):1409–1421.

6. Napel SA, Beaulieu CF, Rodriguez C, Cui J, Xu J, Gupta A, et al. Automated Retrieval of CT Images of Liver Lesions on the Basis of Image Similarity: Method and Preliminary Results. Radiology. 2010 Jan. 1;256(1):243–252.

7. Korenblum D, Rubin D, Napel S, Rodriguez C, Beaulieu C. Managing Biomedical Image Metadata for Search and Retrieval of Similar Images. J Digit Imaging. Springer New York; 2011 Jan. 1;24(4):739–748.

8. Rubin DL, Rodriguez C, Shah P, Beaulieu C. iPad: Semantic annotation and markup of radiological images. AMIA Annu Symp Proc. 2008;:626–630.

9. Strela V, Heller PN, Strang G, Topiwala P, Heil C. The application of multiwavelet filterbanks to image processing. Image Processing, IEEE Transactions on DOI - 10.1109/83.753742. 1999;8(4):548–563.

10. Zhao CG, Cheng HY, Huo YL, Zhuang TG. Liver CT-image retrieval based on Gabor texture. Engineering in Medicine and Biology Society, 2004. IEMBS '04. 26th Annual International Conference of the IEEE DOI - 10.1109/IEMBS.2004.1403458. IEEE; 2004. p. 1491–1494.

11. Hong B-W, Prados E, Soatto S, Vese L. Shape Representation based on Integral Kernels: Application to Image Matching and Segmentation. Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on DOI - 10.1109/CVPR.2006.277. IEEE; 2006. p. 833–840.

12. Manay S, Cremers D, Byung-Woo Hong, Yezzi AJ, Soatto S. Integral Invariants for Shape Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence. Published by the IEEE Computer Society; 2006 Oct. 1;28(10):1602–1618.

13. M Rangayyan R. Biomedical Image Analysis↱. Neuman MR, editor. Boca Raton, Florida: CRC Press; 2005. p. 1272.

14. Xu J, Faruque J, Beaulieu C, Rubin D, Napel S. A Comprehensive Descriptor of Shape: Method and Application to Content-Based Retrieval of Similar Appearing Lesions in Medical Images. J Digit Imaging. Springer New York; 2012 Jan. 1;25(1):121–128.

15. Rosset A, Spadola L, Ratib O. OsiriX: An Open-Source Software for Navigating in Multidimensional DICOM Images. J Digit Imaging. Springer New York; 2004 Jan. 1;17(3):205–216.

16. Langlotz CP. RadLex: a new method for indexing online educational materials. Radiographics. 2006;26(6):1595–1597.

17. MacKay D. Information theory, inference, and learning algorithms. 7th ed. Cambridge: Cambridge University Press; 2003. p. 640.

18. Wang JZ, Wiederhold G, Firschein O, Xin Wei S. Content-based image indexing and searching using Daubechies' wavelets. International Journal on Digital Libraries. Springer Berlin / Heidelberg; 1998 Jan. 25;1(4):311–328.

19. Duda RO, Hart PE. Pattern classification and scene analysis. New York: Wiley: A Wiley-Interscience Publication; 1973.

20. Kilday J, Palmieri F, Fox MD. Classifying mammographic lesions using computerized image analysis. Medical Imaging, IEEE Transactions on DOI - 10.1109/42.251116. 1993;12(4):664–669.

21. Tibshirani RT. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological). Blackwell Publishing for the Royal Statistical Society; 1996 Jan. 1;58(1):267–288.

22. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2009 Dec.;33(1):1–22.