Comparison of Machine Learning Techniques for Magnetic Resonance Image Analysis

Wendy Ni, Xinwei Shi, Umit Yoruk

Department of Electrical Engineering & Department of Radiology, Stanford University

1. Background

Magnetic resonance imaging (MRI) is a powerful non-invasive medical imaging technique that encodes the mechanical, physiological and chemical structure of soft tissues. However, manual segmentation of tissue regions of interest (ROIs) can be a laborious process prone to operator error. In this project, we compared algorithms from 3 classes of supervised machine learning (ML) techniques for MRI segmentation in 3 applications:

1.1 Dynamic Renal Imaging

Quantitative analysis of MR urography data is important for early detection and staging of chronic kidney disease. The patients are injected with a contrast agent and the distribution of the contrast is observed using MRI over a period of 3-4 minutes. The pharmacokinetic model used in renal filtration rate estimation needs concentration vs. time curves of the aorta and renal cortex. In order to extract these curves segmentation must be performed on the MR images.

1.2 Knee Cartilage Imaging

MRI has been an emerging technique in quantitatively assessing morphological and physiological change of the cartilage for evaluation of osteoarthritis (OA). To get quantitative assessment, the cartilage in MR images needs to be precisely labeled.

1.3 Stroke Imaging

MRI is invaluable for assessing brain tissue viability following a stroke. The FLAIR [1] and GRE [2] sequences are commonly used to identify the morphological extent of stroke lesions. For quantitative analysis, the often poorly-delineated lesions need to be segmented slice-by-slice.

2. Data & Features

All datasets were manually segmented by experts as ground truth. The raw images were cropped to remove background area. Each voxel was considered as one sample to be classified. The intensity features were represented by 256 quantiles to remove the inter-subject variances and all features are normalized to [0,1] range.

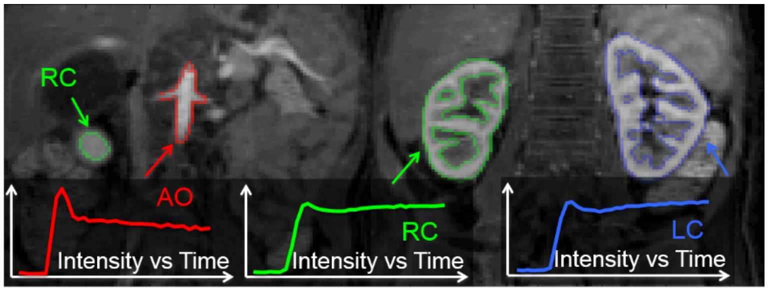

Fig 1. Ground truth for kidney at peak intensity for RC and LC.

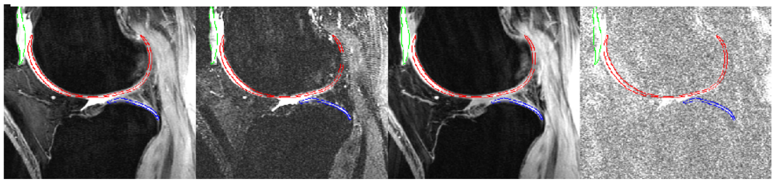

Fig 2. Knee images of 4 contrasts and ground truth (red: femur cartilage, green: patella cartilage, blue: tibia cartilage).

In renal segmentation, 11 sets of abdominal dynamic contrast enhanced (DCE) [3] images (size 192×180×80 cropped to 100×130×50, 36 time points) were used. Label categories were: aorta (AO), right cortex (RC), left cortex (LC). We used 8 features: voxel coordinates, times to 50% and 90% peak intensity, initial, 25% and peak intensity. The rationale behind feature selection is that there is a difference in the enhancement rate of different tissues (Fig. 1). Since the contrast is injected intravenously, we expect the signal from aorta to start rising first and reach the highest intensity before other tissues. As the contrast diffuses, the signals from the other tissues rise as well. However, the signal from the cortex rises faster than the other tissues because the contrast that moves into the kidneys cannot leave (filtration) and starts accumulating.

In cartilage segmentation, 4 sets of knee DESS [4] images (size 512×512×42 cropped to 300×300×11, 4 contrasts with different diffusion weighting and T2 decay) were used (Fig. 2). Label

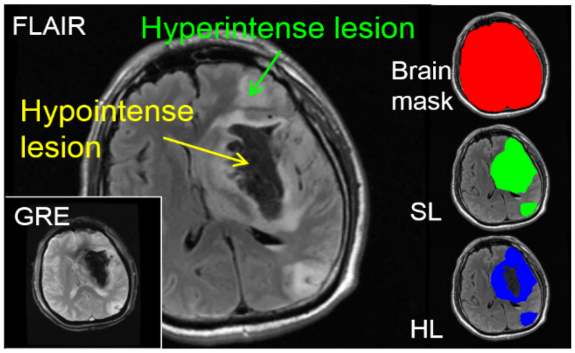

Fig 3. Ground truth for stroke segmentation.

categories were: femur (FE), patella (PA) and tibia (TI). We used 103 features: voxel intensity, coordinates, and derived features related to edges and curvatures [5] including (1) the gradient of the smoothed image; (2) largest eigenvalue and corresponding eigenvector of the Hessian: represents the principal curvature and direction of the curvature; (3) Largest eigenvalue and corresponding eigenvector of structure tensor (ST): represents the strength of the image edges, and the vertical direction of the edge. The rationale behind feature selection is that the cartilage can be modeled as a thin curved disc. The features except coordinates are repeated among the 4 contrasts. For Naïve Bayes, we found the correlation among features would lead to problematic results, and therefore only 13 features from 1 contrast were used in Naïve Bayes.

In stroke lesion segmentation, 10 sets of brain FLAIR and GRE images (sizes 512×512×var and 256×256×var reduced to 256×256×2) were used (Fig. 3). We studied 2 labeling schemes: all stroke lesions (SL) and hyperintense lesions only (HL). We used 44 features including: (1) voxel intensity statistics such as normalized mean and deviation from global mean; (2) radial coordinate; (3) symmetry measures, such as anterior/superior and left/right differences; (4) hemispheric intensity statistics, including ipsilateral hemispheric mean, deviation of voxel intensity from ipsilateral hemispheric mean, and voxel intensity normalized by contralateral mean; (5) edge statistics, including low-pass-filtered brain edge and tissue regional edge; (6) multi-scalar neighborhood intensity statistics including mean and standard deviation; and (7) 1D spatial gradients in anterior/superior and left/right directions.

3. Methods

3.1 Models

We compared 4 models from 3 classes of ML techniques, including Naïve Bayes (NB), Logistic/Softmax Regression (LR) and SVMs with the Gaussian kernel (SVM-G) and the linear kernel (SVM-L). We implemented NB and LR using custom Matlab code. We utilized the LibSVM [6] library for SVM-L and SVM-G, and optimized parameters using grid search.

The cartilage segmentation was treated as a multiclass problem, while the other two applications were two-class problems. Therefore, Softmax Regression and multi-class SVMs with the one-against-one strategy were used for cartilage segmentation; and Logistic Regression and standard two-class SVMs were used for other applications. In renal segmentation, the features were mapped to a higher dimensional space by quantizing each feature into 256 levels and representing each level with a binary feature in SVM with the linear kernel, which significantly improved the performance of SVM-L.

3.2 Validation

Since in all of the applications the classes are unbalanced, we inspected training and test recall and precision. Method comparison used F1-score as metric, which is computed by $\mathrm{F1} = 2\times\frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$.

K-fold cross-validation was used in evaluation, and K was decided based on total number of datasets, number of samples in one dataset, and computation time for different applications. The voxels in each dataset were undersampled for speedup processing. In renal segmentation, K = 11 (training on 10 sets and testing on 1 set), and the training sets were randomly undersampled by a factor of 5 in the case of SVM-L and SVM-G. In cartilage segmentation, K = 4 (training on 3 sets and testing on 1 set), and the data was randomly undersampled by ~7. In stroke lesion segmentation, K = 5 (training on 8 sets and testing on 2 sets), and the data was undersampled by a factor of 20 for speed. The undersampling is deterministic, but because only the brain voxels are selected, sampling at regular intervals still provides good coverage across the brain.

4. Results & Discussion

4.1 Comparison between Techniques

The test precisions and recalls are summarized in Table 1. The training and test F1-scores are summarized in Table 2.

For all applications, we found SVM-G to be the best, with high and balanced precision and recall. SVM-L was comparable in performance, and takes less time to train. Therefore, for best segmentation results we recommend SVM-G; if the computing resources are limited, as the training set size increase it may be more practical to use the SVM-L.

For all applications, Naïve Bayes had high recall but low precision, while Logistic Regression was the opposite. For cartilage segmentation, Naïve Bayes performance was poor due to failure of the assumption of conditional independence of features (detailed discussion in 4.3).

Table 1. The test recalls and precisions of different ML techniques.

		Naive Bayes		Logistic Regression		SVM Linear Kernel		SVM Gaussian Kernel	
		Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.
Renal	Aorta	0.91	0.57	0.62	0.71	0.7	0.8	0.76	0.83
	R. Cortex	0.84	0.66	0.66	0.7	0.73	0.78	0.75	0.78
	L. Cortex	0.73	0.38	0.33	0.43	0.47	0.55	0.63	0.62
Cartilage	Femur C.	0.66	0.21	0.57	0.72	0.76	0.63	0.88	0.65
	Patella C.	0.64	0.24	0.74	0.76	0.9	0.76	0.88	0.75
	Tibia C.	0.45	0.32	0.36	0.74	0.62	0.63	0.69	0.6
Stroke Lesion	S. Lesions	0.73	0.46	0.48	0.74	0.55	0.73	0.62	0.7
	H. Lesions	0.79	0.45	0.53	0.76	0.61	0.75	0.65	0.72

Table 2. The training and test F1-scores of different ML techniques.

		Naive Bayes		Logistic Regression		SVM Linear Kernel		SVM Gauss Kernel	
		Training	Test	Training	Test	Training	Test	Training	Test
Renal	Aorta	0.7	0.69	0.7	0.69	0.91	0.73	0.75	0.77
	R. Cortex	0.8	0.77	0.73	0.68	0.91	0.75	0.92	0.77
	L. Cortex	0.59	0.58	0.48	0.41	0.84	0.53	0.89	0.68
Cartilage	Femur C.	0.36	0.31	0.77	0.64	0.77	0.69	0.8	0.75
	Patella C.	0.43	0.35	0.9	0.75	0.9	0.82	0.88	0.81
	Tibia C.	0.43	0.38	0.56	0.48	0.62	0.62	0.64	0.64
Stroke Lesion	S. Lesions	0.56	0.56	0.73	0.58	0.76	0.62	0.97	0.66
	H. Lesions	0.57	0.58	0.78	0.63	0.81	0.67	0.94	0.68

4.2 Error Analysis

Sample images with SCM-G segmentation results are shown in Fig. 4-6. For all applications, there are three common fundamental limits: (1) partial volume effect (mixing of signals from different tissues) affecting segmentation of fine structures; (2) ground truth prone to subjective factors; (3) system related or subject related inter-scan variations making generalization challenging.

In renal segmentation, segmentation of cortex is more challenging than the aorta. This is partly due to the more complex structure of the kidneys. Cortex consists of many small structures and the large slice thickness of the images cause partial volume effect. This in turn causes labeling inaccuracy in the training set limiting the ML performance. In some cases, parts of the spleen was mislabeled as left cortex because it is in close proximity and has similar signal enhancement to the renal cortex. Finding more discriminative features such as local image statistics could improve the labeling accuracy of left cortex.

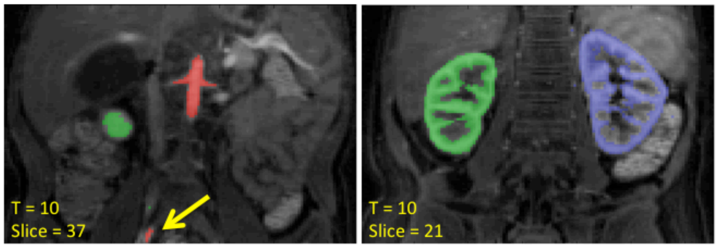

Fig 4. Renal segmentation using SVM-G. Aorta (red), right cortex (green) and left cortex (blue) segmented correctly. The arrow points to an artery, which was missed during manual labeling but correctly classified by SVM. (See Fig. 1. for the ground truth.)

In cartilage segmentation, automatic segmentation did relatively well in patella cartilage, which has relatively homogeneous intensity and simple shape. Femur and tibia cartilages have very inhomogeneous intensities and are more susceptible to partial volume effect, therefore the segmentation is more challenging. The manual labeling of these two parts is prone to subjective factors: comparing the arrows in Fig. 5, the similar region was manually labeled as cartilage in the upper case, but not in the lower one.

In stroke segmentation, the overall performance was good even for highly inhomogeneous and diffuse lesions. Stroke lesion segmentation is very challenging due to aforementioned homogeneity in lesion features, often diffuse appearance of lesions (both due to underlying physiology and partial volume effect), and similarity to healthy tissues in many features. There is also significant inter-subject variation in lesion characteristics and size, indicated by the gap between training and test F1-scores. Therefore, it is very important to have large training sets covering a wide of range of lesions characteristics.

The segmentation results can be improved by using image processing methods, such as graph-cut, to promote smoothness and continuity.

4.3 The Influence of Correlated Features in Naïve Bayes

Naïve Bayes assumes that the features are conditionally independent. When some features are highly correlated, they are effectively given higher "weights" in the classifier.

In the cartilage segmentation problem, some features can be highly correlated since they are repeated for different contrasts, or reflect similar characteristics. When all 103 features were used, NB had very poor precision (~0.1) and misclassified tissues with similar intensity as cartilage, as a result of heavy weighting on intensity-related features. After removing similar features, the performance of NB was improved, but still worse than other ML techniques. This can be explained by that the correlation coefficient is still large between some features, such as intensity and Hessian eigenvalues, as shown in Fig 7.

The results suggest that the independence assumption of NB is hard to satisfy when complex features are used. Therefore NB is not a good fit for the cartilage segmentation problem.

4.4 Feature Dimension Reduction

In cartilage segmentation problem, there is redundancy in the 103 dimensional features, and the training of SVM takes a long time due to the high dimensionality of features. PCA and feature selection by filtering were incorporated to reduce feature dimension. For PCA, the threshold was set as 0.1% of sum of eigenvalues, and the feature dimension was reduced to 64. For feature selection by filtering, the F-score [7] was used as metric of discrimination of features, and the top 56 features were selected. As shown in Table 3, the feature selection by F-score filtering was better in terms of training time reduction and classification performance. The training time is reduced by half from 19 minutes, while the

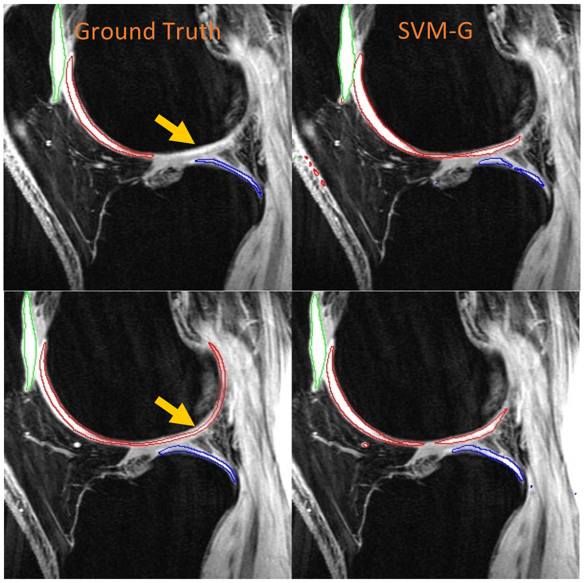

Fig 5. Cartilage segmentation using SVM-G: femur cart(red), patella cart(green), tibia cart(blue). The two rows are two different datasets.

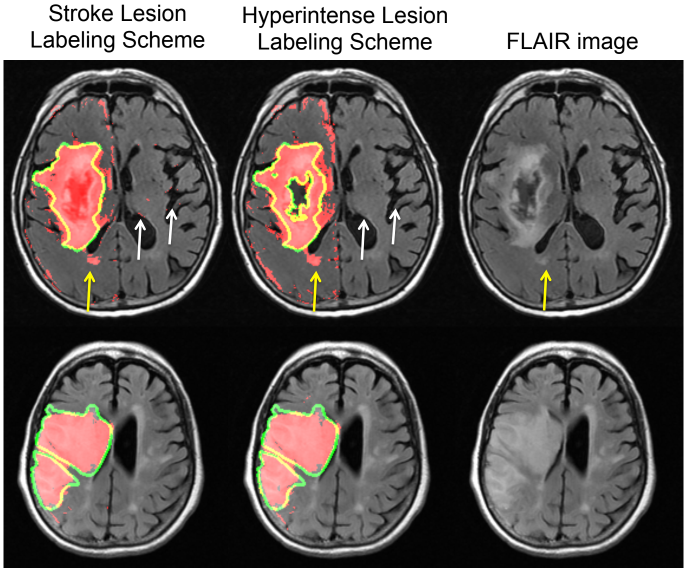
Stroke Lesion Labeling Scheme
Hyperintense Lesion Labeling Scheme
FLAIR image

Fig 6. Stroke lesion segmentation using SVM-G. The top row is a patient with a hematoma (dark lesion), which is delineated well in the hyperintense lesion labeling scheme. White arrows indicate isolated voxels that are incorrectly identified when hypointense voxels are not excluded from the model. Yellow arrows indicate a region of FLAIR hyperintensity that is classified by experts as normal tissue but is incorrectly identified as lesion. The bottom row is a patient without a hematoma and with very good segmentation results.

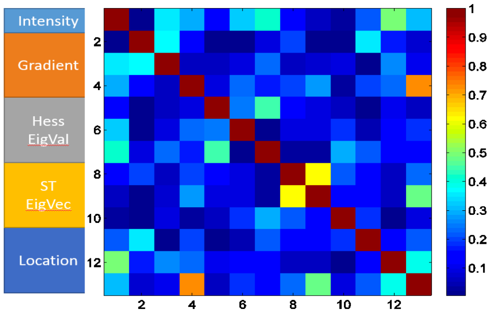

Fig 7. Cross correlation matrix of 13 features for femur cartilage.

precision, recall and F1-score of the results is not significantly affected.

4.5 Training/Test Data Undersampling

To confirm that undersampling did not significantly affect the performance of ML, we inspected the training and test F1-scores versus number of voxels per dataset. We found that the saving in training time justified the small reduction in F1-scores at undersampling factor 20, as shown in Fig 8.

4.6 Training Set Size vs F1-Score

We analyzed the effect of training set size on the F1-scores using SVM-G. The results in renal segmentation application are shown in Fig 9. In other applications, similar trends were observed: as the training set size increased, the training F1-scores decreased and the test F1-scores increased. In the segmentation of aorta, the training and test F1-scores fully converged after including 8 datasets. The training and test F1-scores of the renal cortex had a converging behavior but they didn't fully converge after including 10 kidneys. This can be explained by that the kidneys vary significantly in shape and size as well as the intensity changes (e.g. fully functional kidney takes up the contrast faster than a diseased kidney). In addition, some of the patients only had one kidney and this effectively reduced the training set size. In fact, the training and test F1-scores of the right cortex (present in 10, missing in 1 patients) converged better than the left cortex (present in 8, missing in 3 patients).

5. Conclusion

In this project, we found that for automated segmentation of renal structure, knee cartilage and stroke lesions, SVM with the Gaussian kernel works the best, with high and balanced precision and recall. SVM with the linear kernel is comparable and may be more practical to implement.

6. Future Work

As indicated by Table 2 and Fig 9, the generalization ability of the ML techniques is still confined by limited training datasets. We will expand training set to reduce the effect of inter-patient variation. We will add new features to reduce bias of the models, including local image statistics, dynamic information for renal segmentation, tissue texture characteristics for cartilage segmentation and inter-slice information for stroke lesion segmentation. We will incorporate SVM-G with graph-cut to refine segmentation results. Finally, we will improve ground truths by using multiple experts manually labeling.

	All features	PCA	Feature Selection
Dimensionality	112	64	56
Training time (min)	19	16.3	9.8
Test time (min)	5.4	3.2	2.3
Average F1-score	0.74	0.71	0.73

Table 3. Comparison of feature dimension reduction methods

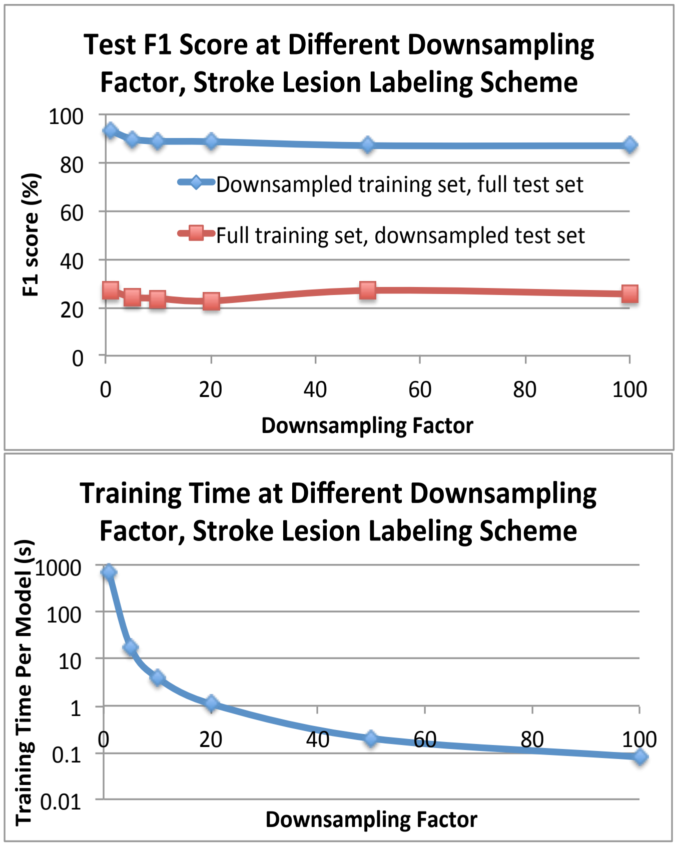

Fig 8. Representative test F1-score and training time curves at different downsampling factors, demonstrating that acceptable error is introduced by downsampling the training data set and test data set, at significant training time savings.

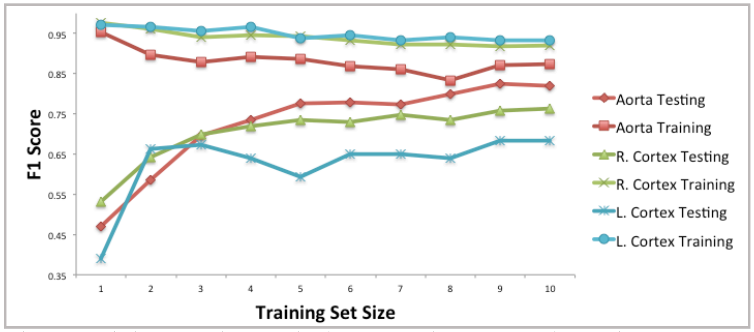

Fig 9. Training set size analysis on renal segmentation using SVM-G.

Reference

[1] Brant-Zawadzki M, et al. Stroke 27.7, 1996. [2] Warach S, et al. Neurology 42.9, 1992. [3] Tofts PS, et al. J Mag Res Im 10.3, 1999. [4] Eckstein F, et al. Ann Rh Dis 65.4, 2006. [5] Folkesson J, et al. MICCAI, 327-334, 2005. [6] Chang C, Lin C. ACM TIST, 2:27:1-27:27, 2011. [7] Chen YW, Lin CJ. Feature extraction, 315-324, 2006.