Lab 4-Binary Classifier
Stat 215A, Fall 2014

Xiang (Lisha) Li, Jonathan Fischer, Andrew Do, Hye Soo Choi

November 10, 2014

1   Introduction

Blah blah...

2   EDA

2.1   Densities of NDAI, SD and CORR

The following three plots gives us a sense of what can be learned from NDAI, SD and CORR. The densities are grouped by their expert labels, red is for 'no cloud', green are for 'unknown' and blue for 'cloud'. We can see that NDAI has reasonably good separation between cloud and no cloud, in all three pictures, which is confirmed in our later modeling sections by Gini importance measures found in the random forest model and it's importance as a variable in LDA/QDA and logit models. SD does not have such a good separation within the lower regions of the SD values, however it is still clear that pixels labelled as clouds are the only values with higher SD values. We thus expect the two features in combination can help determine whether a pixel with high NDAI should be labelled as a cloud by using the SD feature. Finally CORR values appear to be a good separator for image 2, but much less so for image 1. This uneven distribution of CORR values between images prompted us to cross validate our models across images, in addition to the folds created by dividing each image into 4 quadrants.
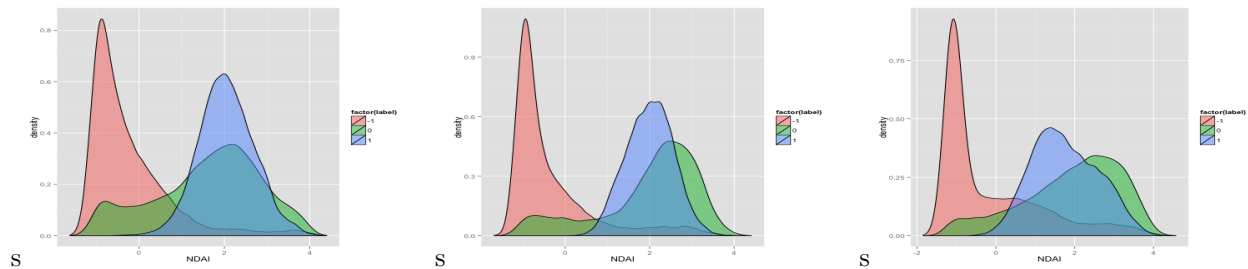

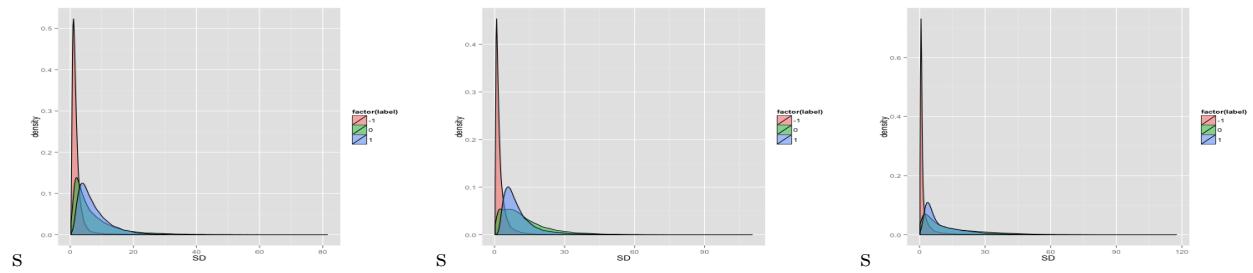Figure 1: NDAI density plot for Image 1, 2, 3 (respectively).

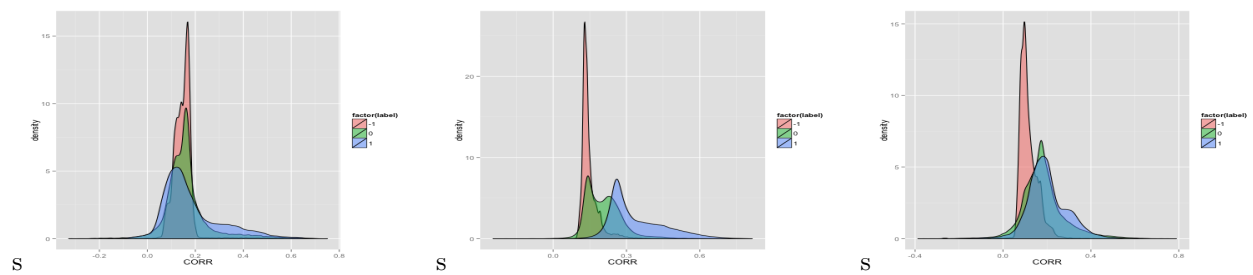Figure 2: SD density plot for Image 1, 2, 3 (respectively).

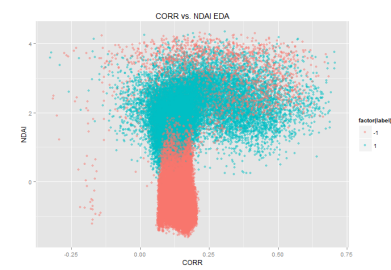Figure 3: CORR density plot for Image 1, 2, 3 (respectively).
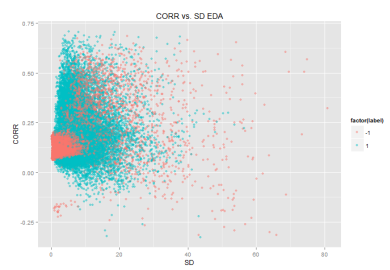

Figure 4: CORR vs. NDAI Plot of Image 1

Figure 5: CORR vs. SD Plot of Image 1
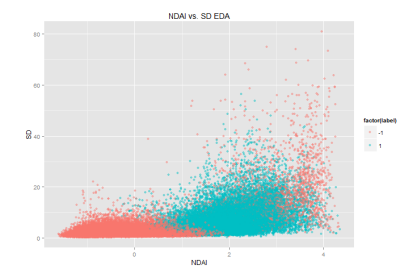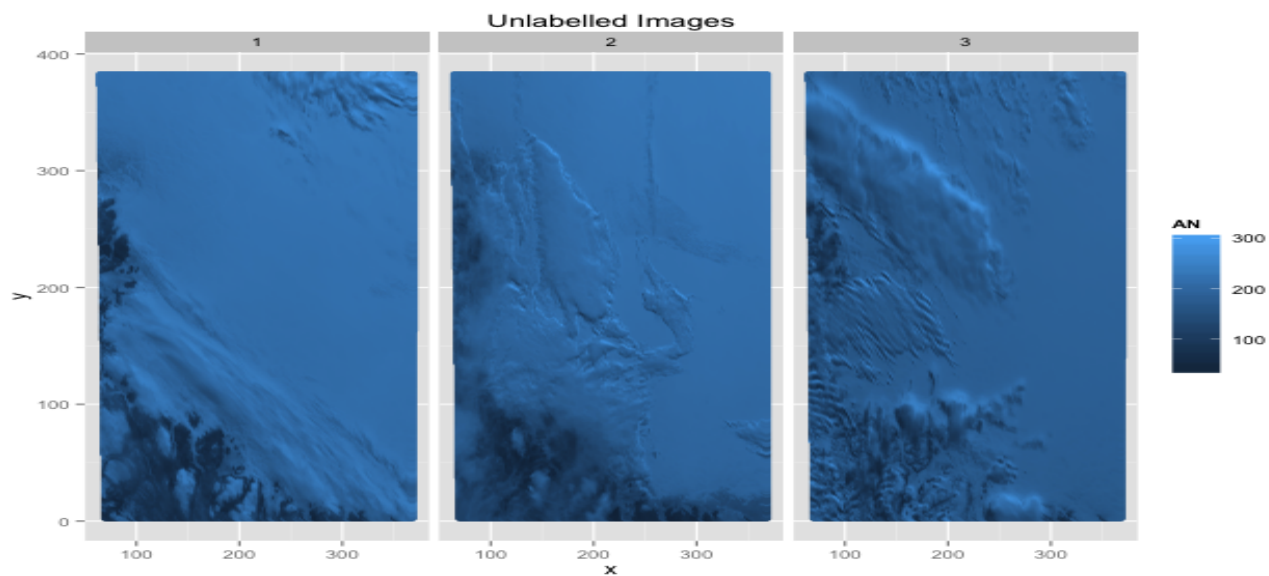
Figure 6: NDAI vs. SD Plot of Image1
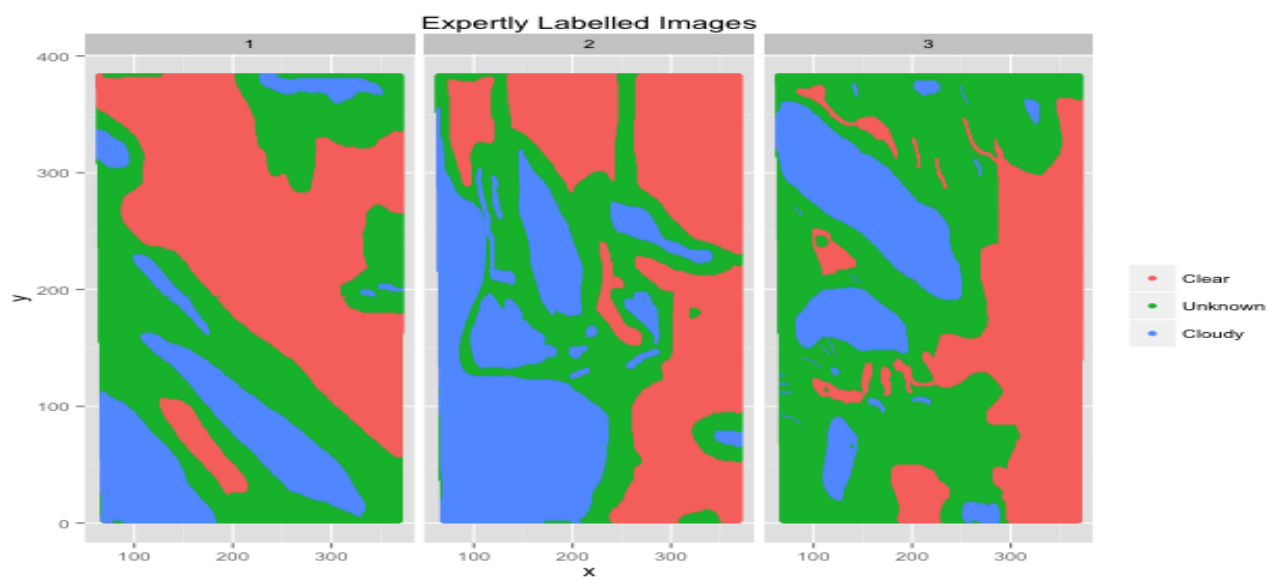
Figure 7: Raw images with AN radiances.

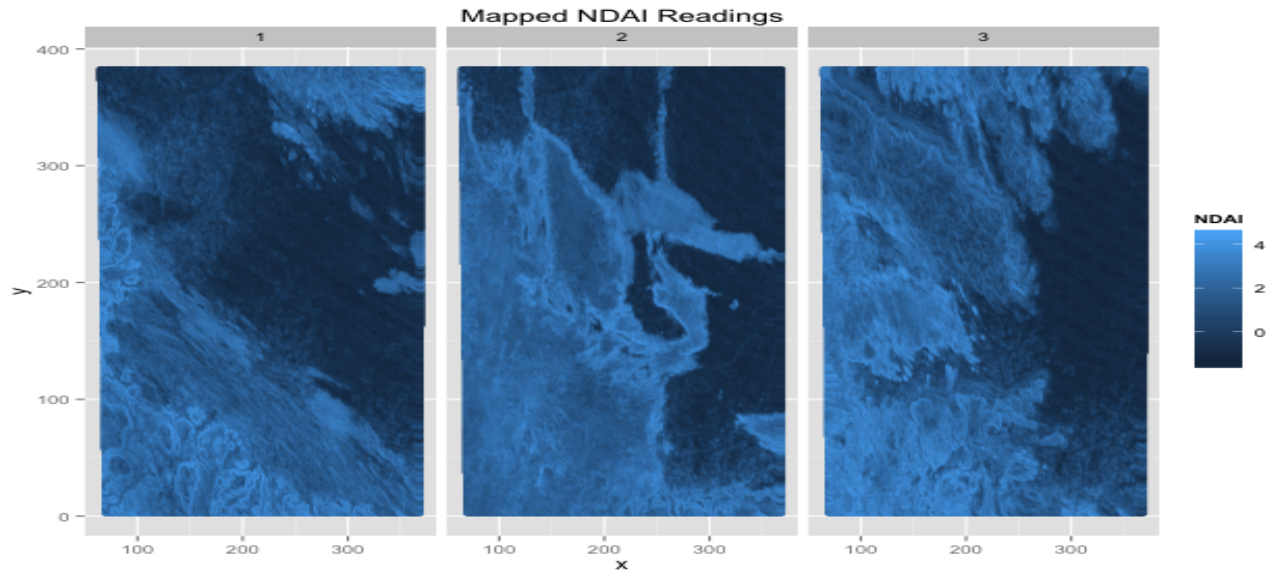Figure 8: Images with expert classifications.

Figure 9: Mapped NDAI readings. We see good correspondence between larger values and presence of clouds.
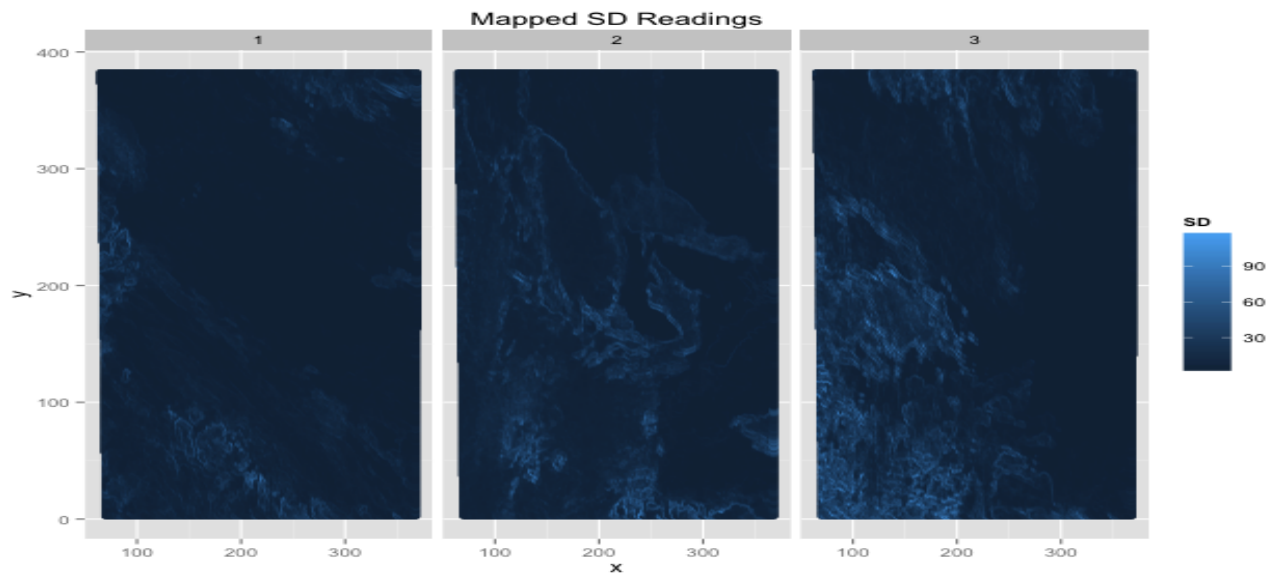

Figure 10: Mapped SD readings. Higher values show cloud boundaries, though also shows uneven terrain.
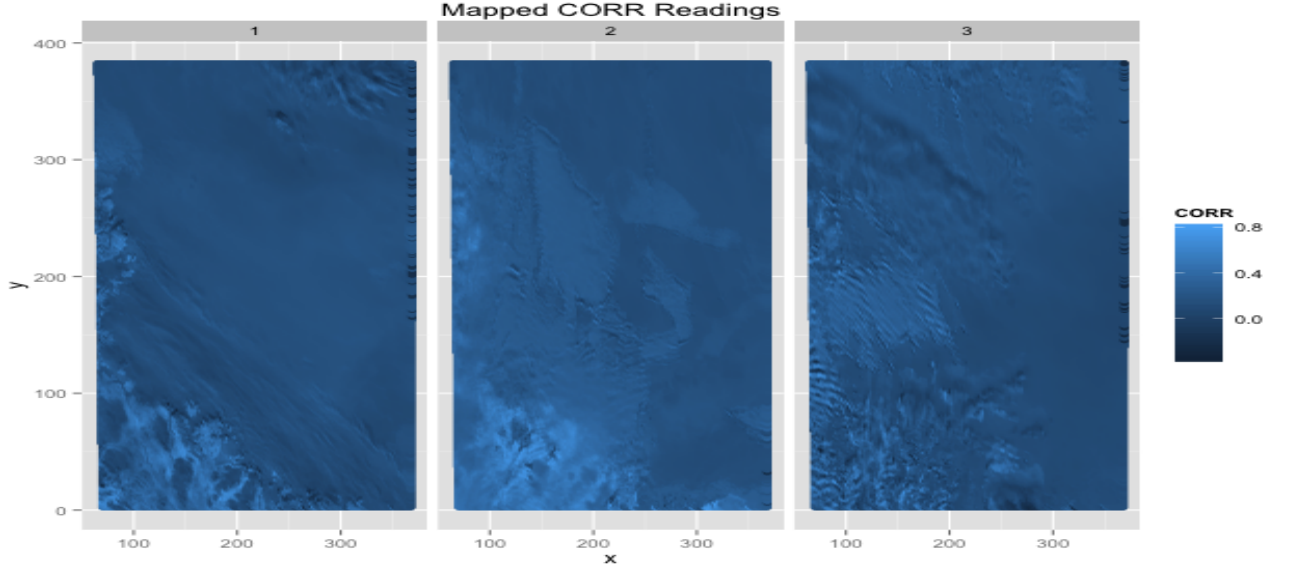
Figure 11: Mapped CORR readings. Cloudy regions tend to be lighter, but not as strongly as in NDAI.

3   Modeling
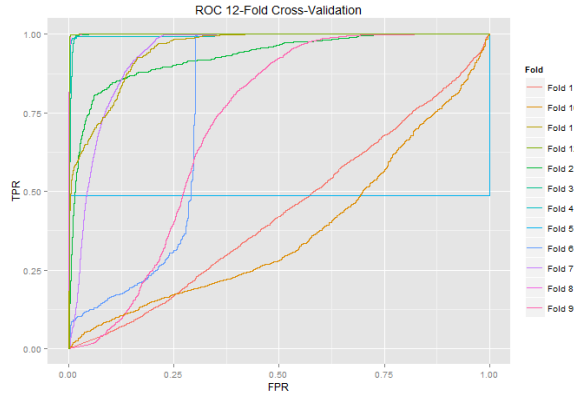
3.1   LDA

3.2   QDA

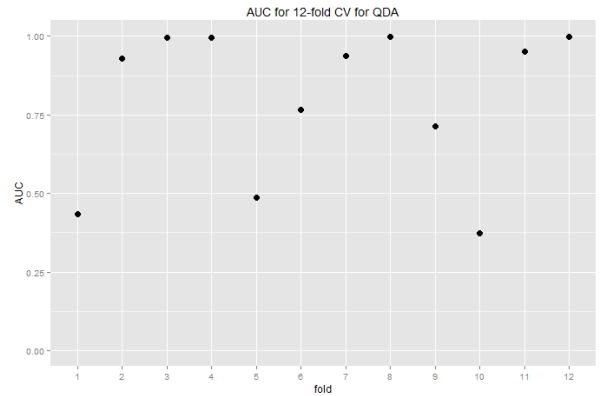Figure 12:   Response Operator Curve for 12-fold
Cross-valudation of QDA

Figure 13: AUC for QDA Classifiers

Cross-validation for QDA revealed that while the method works extremely well in some cases, producing
AUC scores of almost 1, it sometimes fails to perform better than even the theoretical random classifier. In
particular, the classifier does not seem to be good at discerning snow from clouds in regions where there are
many dark pixels. For example, during cross-validation, the watery bottom-left quadrant of image 2 and
ridge-ridden left edge of image 3 poses significant problems to our classifiers.

5

3.3   Logit/Probit

The logit model obtained via averaging after 12-fold cross-validation has $y_i = -3.356 + 1.900 * NDAI_i - 0.074 * SD_i + 9.002 * CORR_i$ where $P_i(\text{Cloud}) = \frac{1}{1+e^{-y_i}}$.
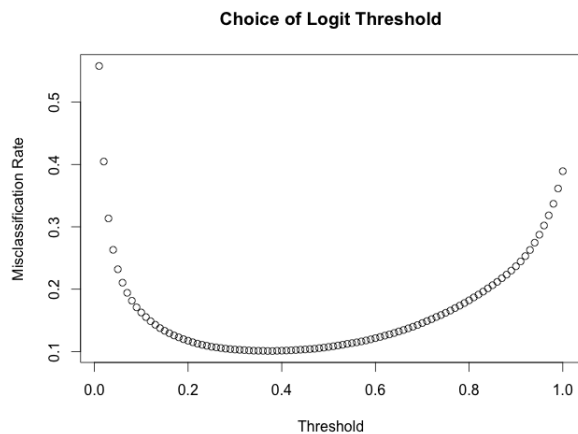

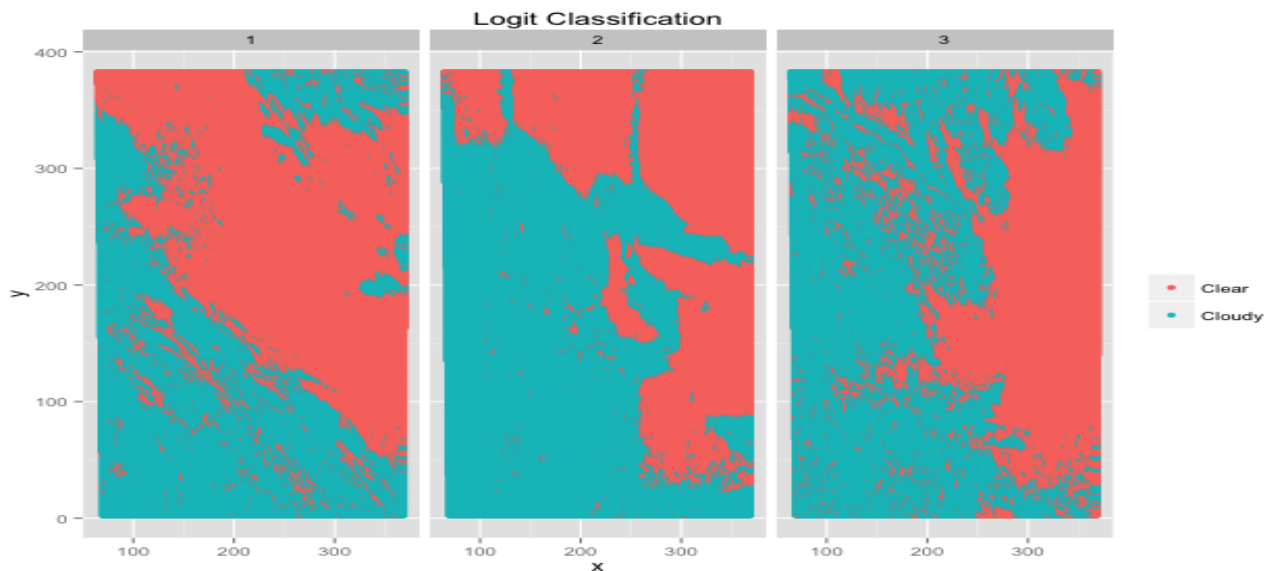Figure 14: We see a minimum in misclassification rate at .38.

Figure 15: Binary classifications with threshold = .38 for logit trained via 12-fold CV.
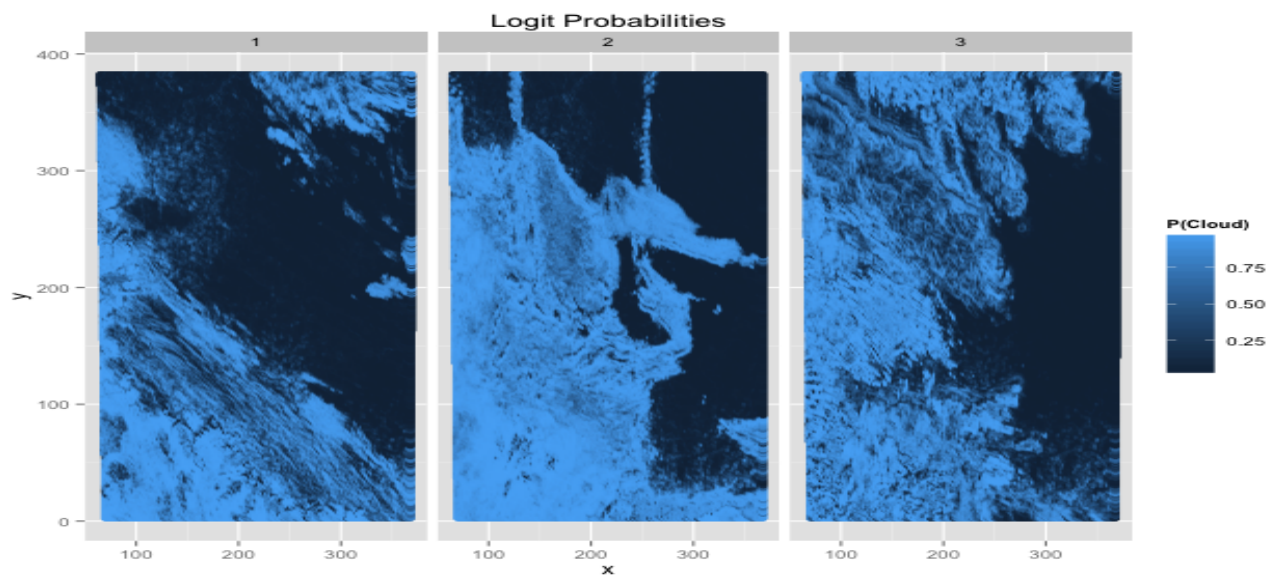
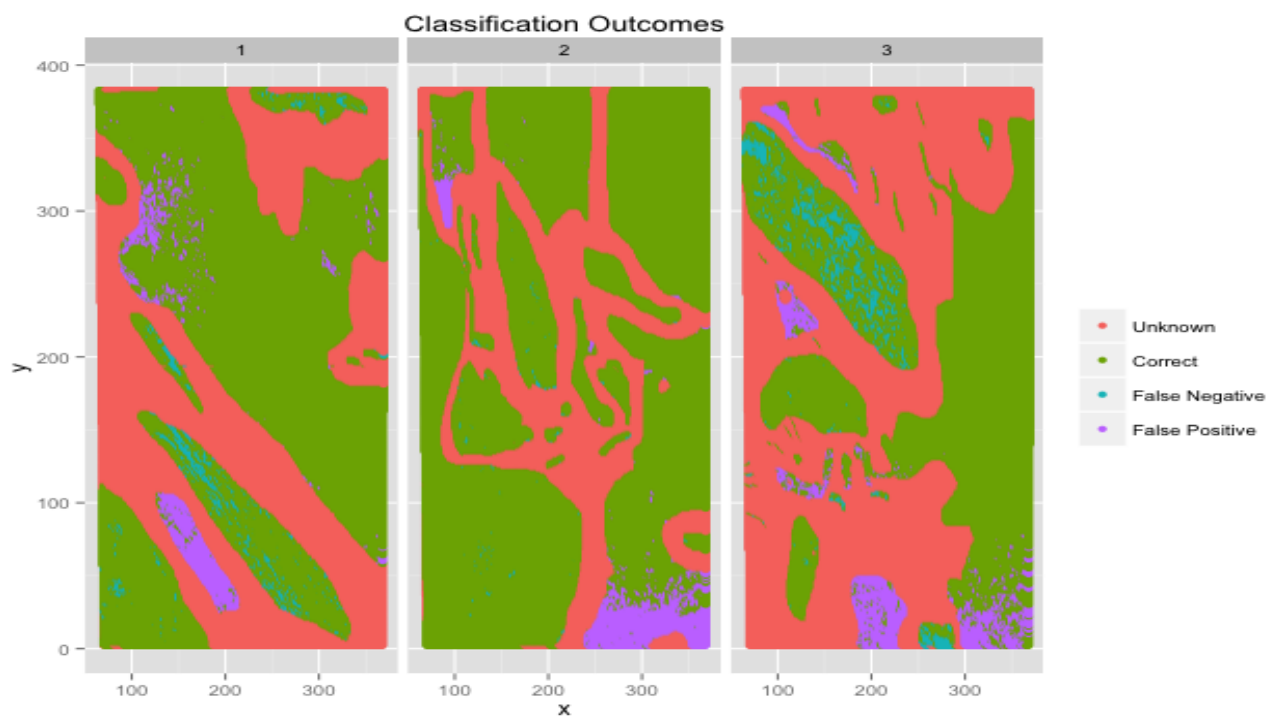Figure 16: Probabilities for logit trained via 12-fold CV.

Figure 17: Classification results from logit with respect to the expert labels.
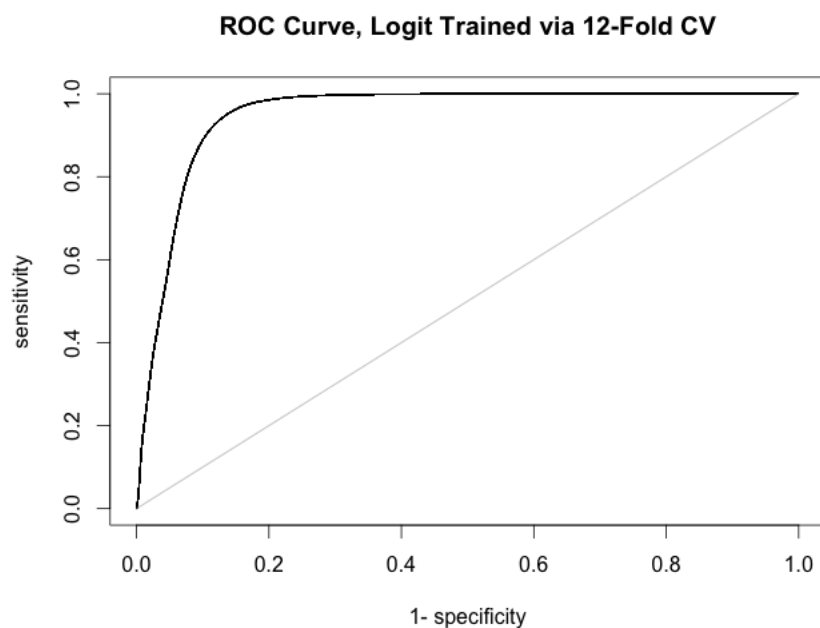
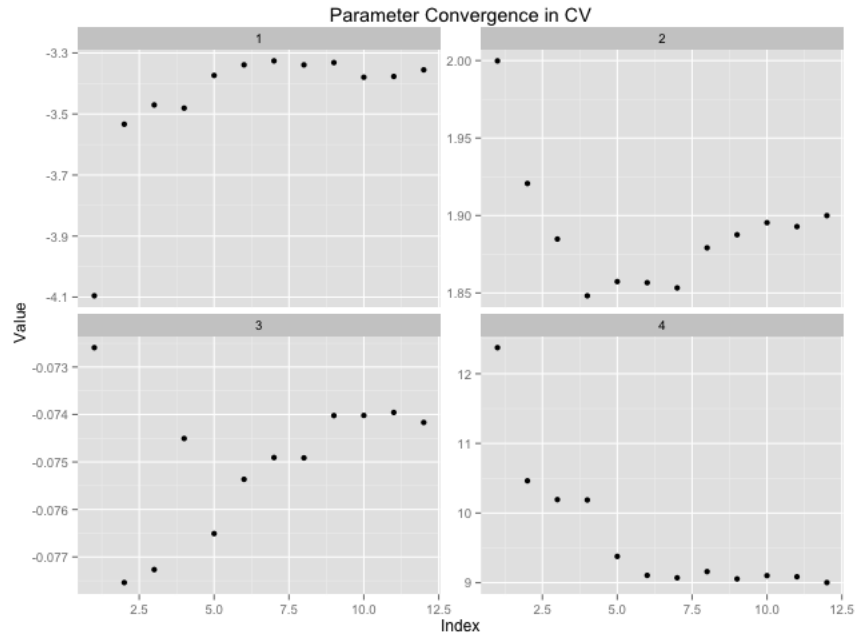Figure 18: ROC curve for our Logit model. AUC is .95.

Figure 19: Coefficient values through iterations of 12-fold CV.

3.4   Random Forest

For random forest, as in all the other classifiers, we divided the three images into equal sized quadrants (2X2) rectangles in order to do 12 fold validation on the dataset. That is, for each iteration of the validation, we dropped one of the quadrants as a test set, and trained on the remaining 11 quadrants. Keeping the images segments disjoint and continuous ensured that our models were picking up on 'higher' level structure of the dataset, and not the continuous variation of neighbouring pixels. We also trained on each image and tested on the remaining two. To test convergence, one of the things we did was increase the training set from including 1 quadrant, to including 2 quadrants, up to including 11 quadrants (using the complement as the test set). For each of the 3 aforementioned classes of training, we trained on a range of forest sizes, from 2 trees to 50. Here are the results:

Finally, this entire set of training models was done with all the features, and then restricted to only SD, CORR and NDAI. These three were particularly chosen because their GiniImportance was consistently ranked at least 2fold above the next highest in all the cross validations. As we also saw with the random forest model that just used SD, CORR, NDAI, it did not fare poorly compared to training the forests on all 9 predictors.
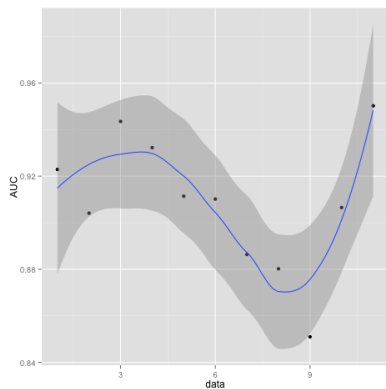

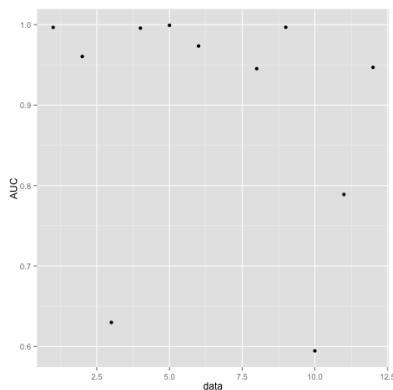Figure 20: Smoothed convergence of AUC for growing training set 50 trees and 3 features

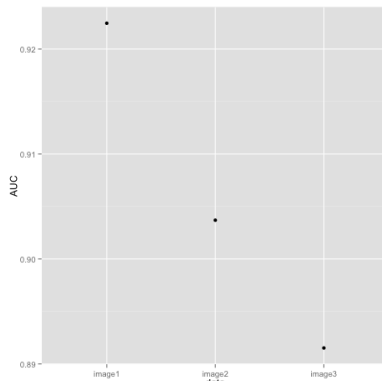Figure 21: AUC of test set in each fold with 50 trees and 3 features

Figure 22: AUC of the three images with 50 trees and 3 features
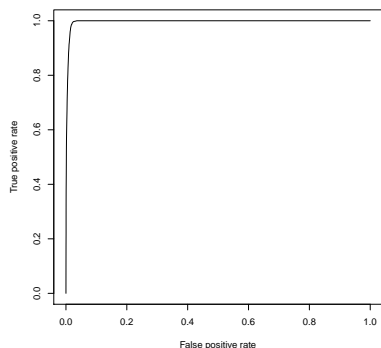
Cross Validation ROC curves
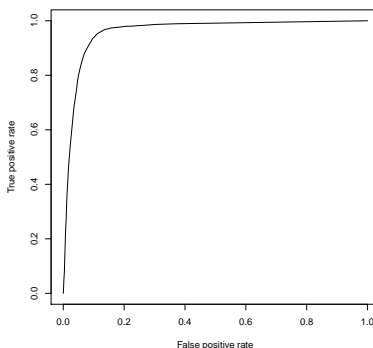

Figure 23: ROC fold 1
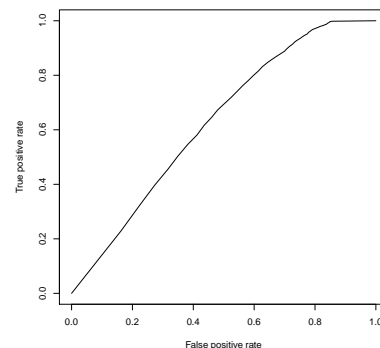

Figure 24: ROC fold 2

Figure 25: ROC fold 3

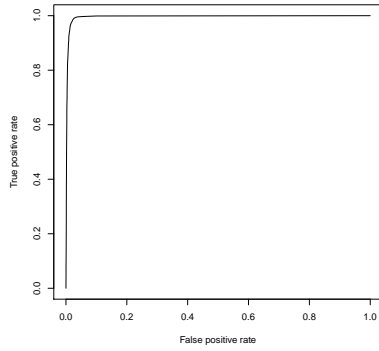**ROC curves for cross validation between images**
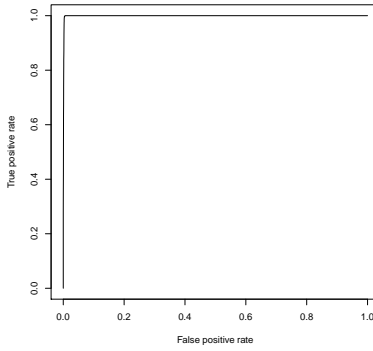
9

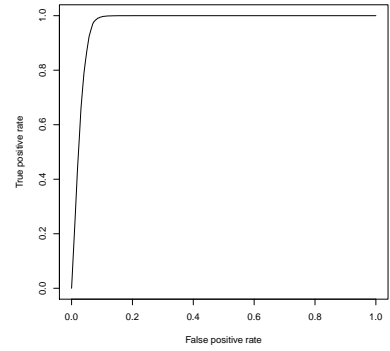Figure 26: ROC fold 4

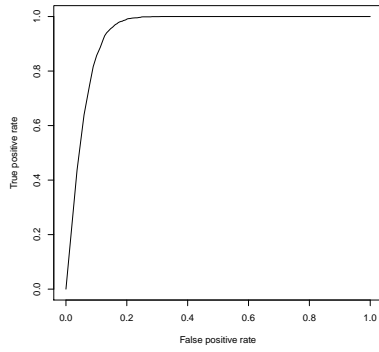Figure 27: ROC fold 5

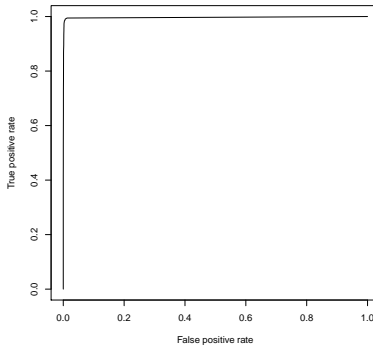Figure 28: ROC fold 6

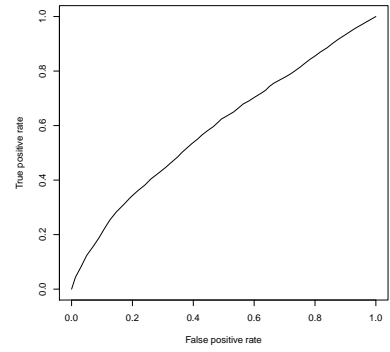Figure 29: ROC fold 8

Figure 30: ROC fold 9

Figure 31: ROC fold 10

4 Reproducibility

How we organized out code and github repo

References

[1] Ben-Hur, A., Elisseeff, A., Guyon, I.: A stability based method for discovering structure in clustered data. In: Pacific Symposium on Biocomputing, pp. 617.
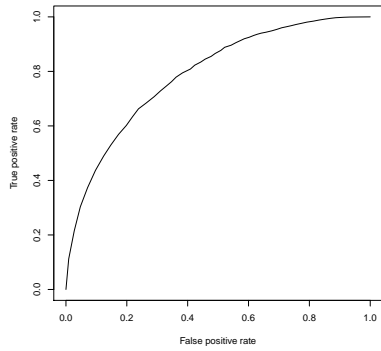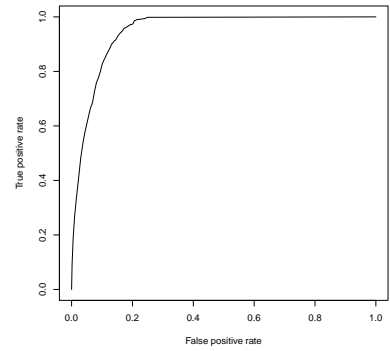
Figure 32: ROC fold 11
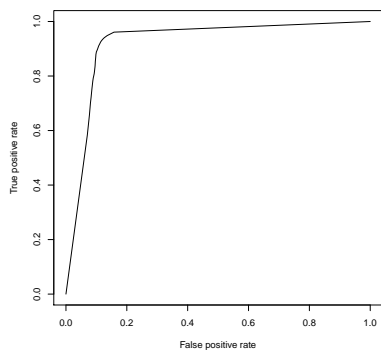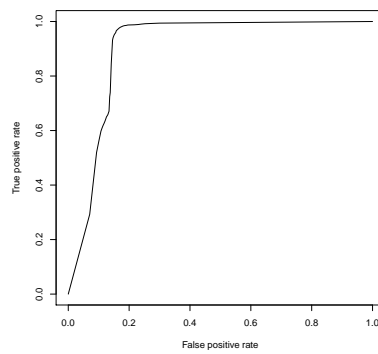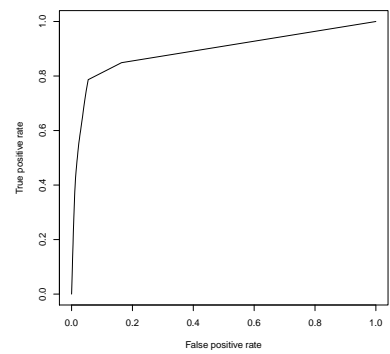

Figure 33: ROC fold 12

Figure 34: Trained on image1

Figure 35: Trained on image2

Figure 36: Trained on image3