CIS 537: Biomedical Image Analysis: Final Project Report, Fall 2018 Computational Dermatology - Developing and Testing Algorithms to Segment Images Based on Hair Density

Siddhi Velankar, Wenbo Zhang, Tejas Shivanand Mane

December 19, 2018

1 Introduction

Alopecia areata (AA) is a "nonscarring hair loss condition"[1]. According to the data from National Alopecia Areata Foundation[2], there are 6.8 million people in the US and 147 million people worldwide having this medical condition. Besides, AA has a total incidence of 20.2 per 100,000 person-years[3]. Thus, quantifying abnormal hair density in AA is needed to track hair loss progression.

Frequently, physicians use simple diameter measurement methods to quantify the scalp region and progression of AA. Furthermore, there are a couple of standardized scoring systems[4] used such as the SALT (Severity of Alopecia Tool) score. Evaluation of the SALT score can be slow due to manual estimation and can be subjective. Hence, quick, reliable and consistent quantification of alopecia areata would be highly convenient for dermatologists, using a computer-based automated segmentation tool. Our intent is to develop such a tool.

Our aims are as follows:

- Implement basic texture analysis algorithms, scalp vs. non-scalp image segmentation to learn about local image cues and gain familiarity with the dataset.
- Further exploration of image features in order to find new ways of encoding hair density and to attempt segmentation of hair into abnormal and normal hair density.
- Combine all our cues, segmentation framework and quantitative analysis of the methods implemented to reach an understanding on the best models and its associated hyperparameters and to segment the images automatically.

2 Dataset

The dataset obtained from CHOP contains 251 unidentified images obtained from 100 patients. For an individual, these images are taken from four different viewpoints: back, right, top and left (Figure 1)

Figure 1: Images of a patient's head in four different orientations

The dataset also contains labelled masks for all these images which categorize every pixel into regions of normal and abnormal hair density. The images in the dataset are diverse due to range of colors, shadow and illumination which makes segmentation challenging. (Figure 2).

Figure 2: Variety in dataset and annotated labels

3 Feature Exploration

To classify a pixel, numerous features, such as color, edge, point, patch size etc can be considered. Selection and exploration of these features is of utmost importance to get accurate segmentation. The basic two sets of features are the intensities of the three colour color channels (RGB) and grayscale intensities. Figure 3 shows the results after converting original image into grayscale and RGB images.

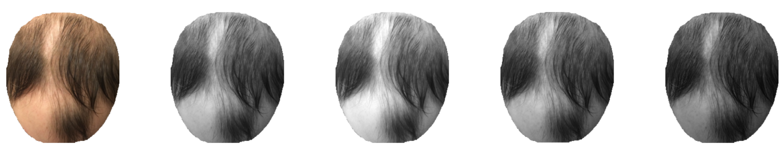

Figure 3: Converted images (original, grayscale, red, green, blue)

A single pixel can only tell us the intensity value. However, a patch contains neighborhood information which could act as a better feature set. We played with the patch size and also the neighborhood statistics (mean value, standard deviation, maximum/minimum value and intensity range), in order to find the best feature set for our problem.

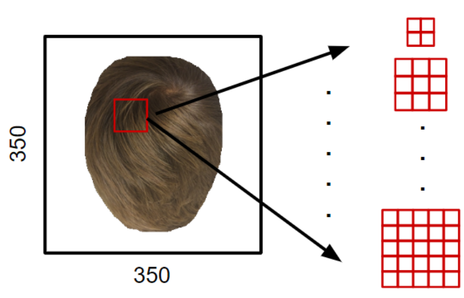

Figure 4: Changing patch size on the image

Additionally, we applied a filter bank (Figure 5) to the image and chose the maximum filter response as each pixel's feature. Because there are six elongated filters with different orientations in filter bank, the filter response (Figure 5) indicates hair line orientation after knowing the maximum filter response.

Figure 5: Filter bank and filter responses

4 Segmentation Methods

We were provided with the dataset containing images of different unidentified patients along with the their respective labeled masks. Since we had labels for our data, the segmentation problem could be treated as an unsupervised or supervised classification problem. Because of the balanced classification dataset, the normal accuracy metric was found to suffice the quantitative evaluation of all our machine learning models. The segmentation accuracy is calculated as follows:

$$\text{Segmentation Accuracy} = \frac{\text{\# of correctly labelled pixels}}{\text{\# of all pixels}}$$

We used both supervised and unsupervised machine learning algorithms perform segmentation on our dataset.

4.1 Unsupervised methods

The unsupervised methods include histogram thresholding and K-means.

Histogram and Thresholding

Using histograms of intensities for grayscale and RGB images, we set the threshold to segment the image into two regions: normal and abnormal hair density.

K-means

We applied K-means using patch information as features. Additionally, the K-means algorithm can be applied on a image individually or on a set of images simultaneously. After applying K-means to the entire dataset, we expect the segmentation accuracy to improve. It can be seen the results section.

4.2 Supervised Methods

Since each image contains 350×350 pixels/data points, 23 training images were found to be sufficient for training. Thus, 23 manually selected images were chosen as the training data and the remaining images were chosen as the test data.

The following machine learning models have been implemented for our problem:

1. K-Nearest-Neighbor (KNN)

2. Random Forest (RF)

3. Naive Bayes (NB)

4. Logistic Regression (LG)

5. Fully Connected/Feed Forward Neural Network (FCNN).

The hyperparameters for these models were tuned to achieve best possible accuracy. For the sake of brevity, further details can be inferred from the provided code.

5 Results and Analysis

The Results for histogram thresholding is shown in Figure 6 and 7. It is usually difficult to find a threshold when only one peak is observed (Figure 6). Besides, even if we do find one, this method is not scalable. Also, in the case of histograms for RGB intensities, selecting three different threshold values for every image is a tough task. (Figure 7)

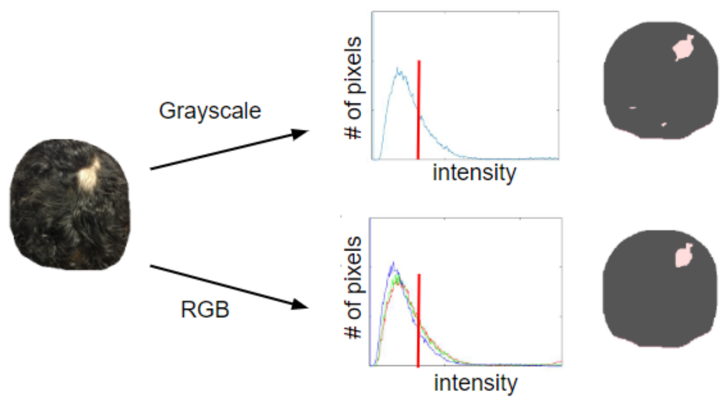

Figure 6: Sample image histogram with one peak

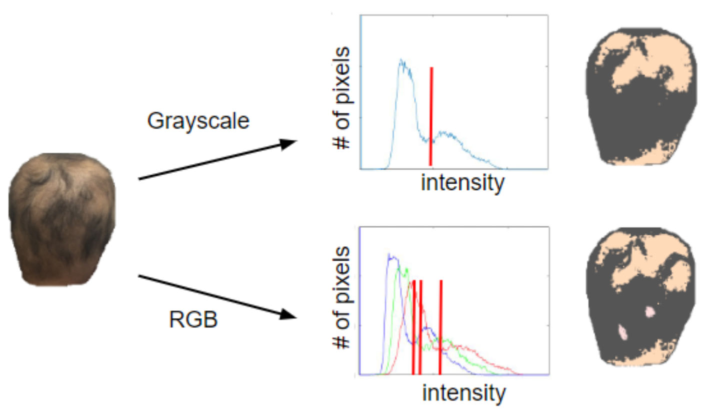

Figure 7: Sample image histogram having difficulty in RGB thresholding

Figure 8 shows the segmentation results using K-means. The first column is the original input image, the second is the ground truth. The next 3 columns represent the results obtained when K-means was performed on images individually using feature set as shown in the Figure 8. The last three images are a result of applying K-means on the entire dataset.

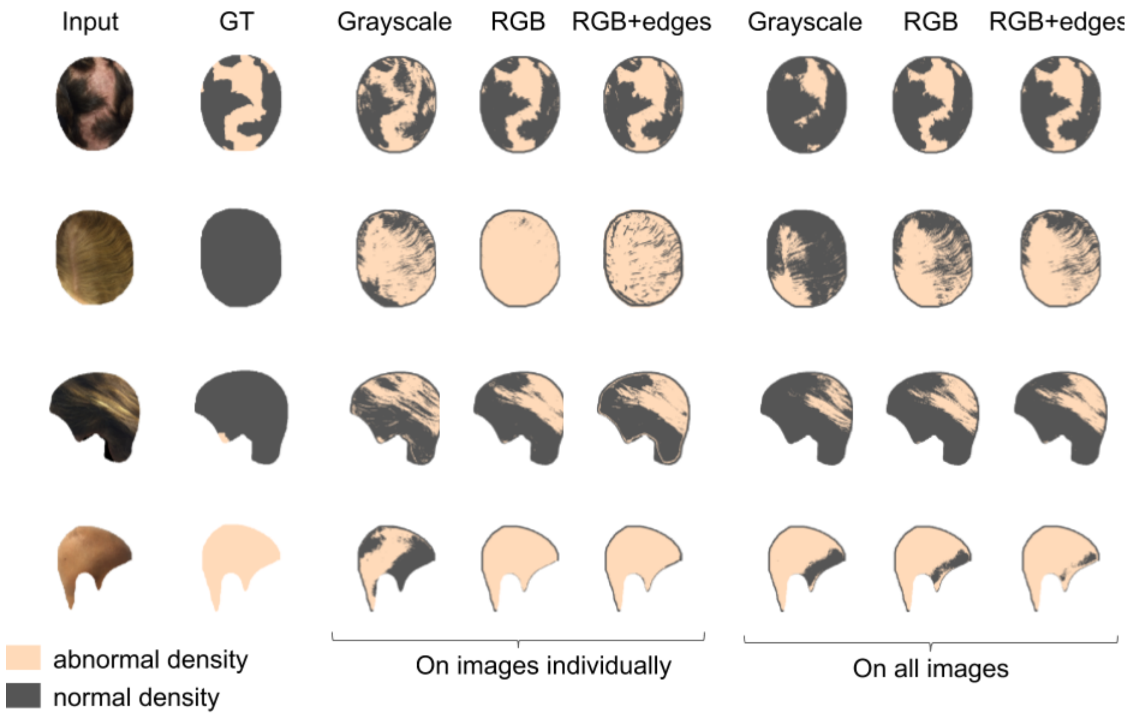

Figure 8: Sample segmentations by K-means

The segmentation accuracies and their respective standard errors for these 6 methods are shown in Figure 9. We shall observe that performing K-means on all the images using RGB as features gives the best accuracy of 77%.

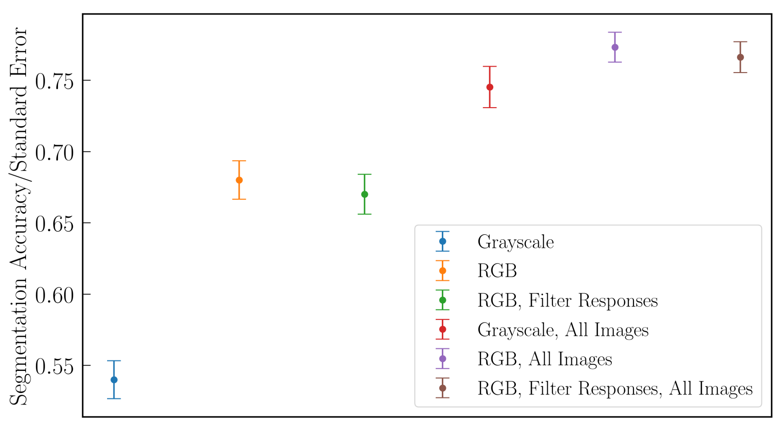

Figure 9: Quantitative analysis for K-means

In Figure 10, we see the segmentation results after applying all machine learning models.

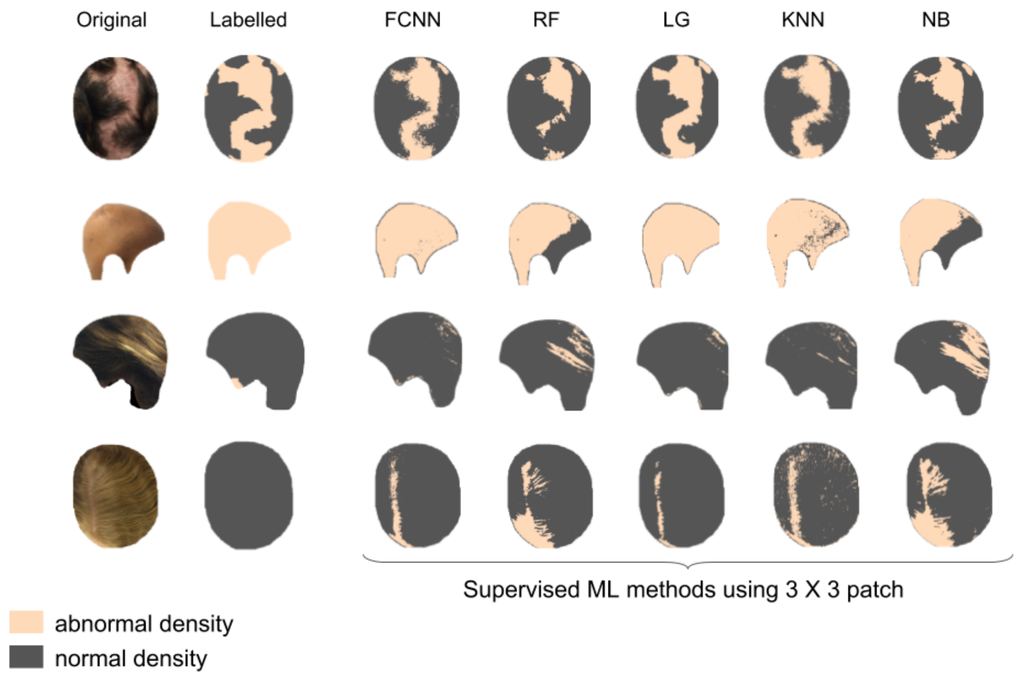

Figure 10: Sample segmentations by ML algorithms

The segmentation accuracy along with standard errors on test dataset for all the machine learning models are shown in Figure 11 and Table 1. We see that a fully connected neural network (FCNN) gives the best accuracy of 83% amongst all the machine learning models. This is around 10% better than the accuracy of "Appopecia" application, which is currently used at the CHOP's clinic.

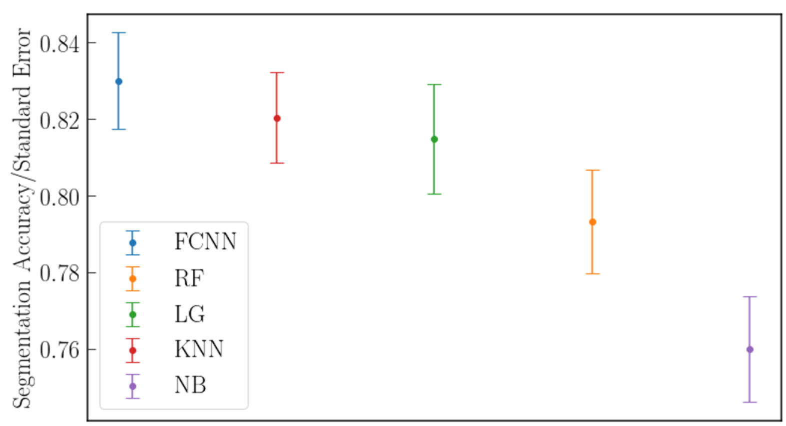

Figure 11: Quantitative analysis for ML algorithms

Table 1: Segmentation accuracy comparison for ML algorithms

Models	FCNN	KNN	LG	RF	NB
Segmentation Accuracy	0.83	0.82	0.815	0.79	0.76

6 Conclusion and Further Improvement

Overall, this project gave us an enriching research experience in biomedical image analysis. Regarding our aims in revised progress report, we believe that we were able to meet them. Furthermore, we were able to achieve a segmentation accuracy of 83% on the test dataset. The conclusions are as follows:

- RGB patches as a feature set is better than grayscale patches.
- Filter responses of elongated Gaussians do not improve segmentation accuracy.
- Supervised ML models perform significantly better than the unsupervised ones.
- There can be visual/clinical disparity associated with meaning of abnormal hair density regions.

With the best accuracy being 83%, there is further scope for improvement. A few cases, where our best model fail, are shown in Figure 12. The first one has white hair and the scalp shows a shadowing effect. The hair is lighter than the scalp and our model incorrectly classifies it as completely scalp. Additionally, the region is also labeled as region of abnormal hair density. The second image has hair, but since it is entirely abnormal, the segmentation turns out to be wrong. In case of the third image, the calculated segmentation accuracy is also very low. Clinically speaking, continuous regions containing hair nodes are considered to be normal. Hence, there is disparity between the labeled mask and our segmentation though the entire image appears to be scalp visually. Hence, further work can be done to improve the labeling of such images to improve our method's accuracy. Additionally, the hyperparameters for the machine learning models could perhaps further be tuned to improve our segmentation.

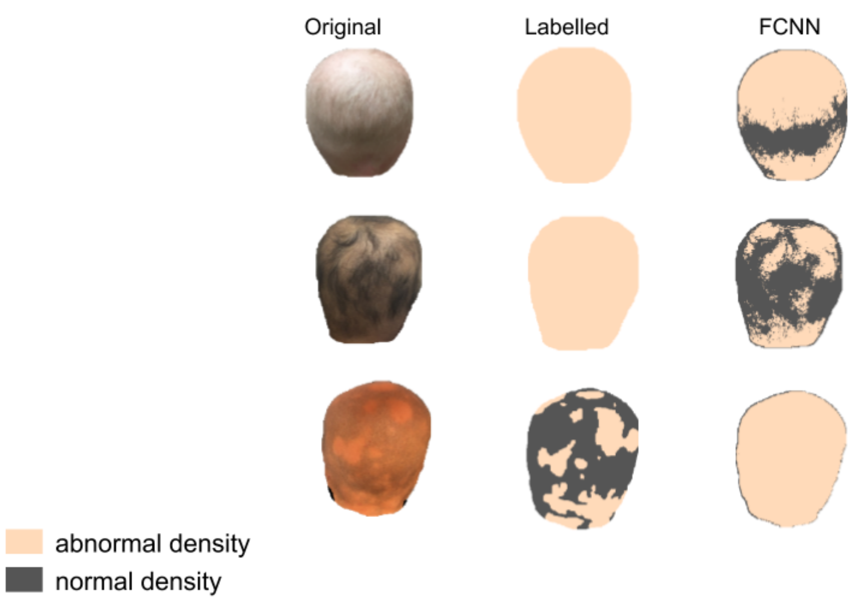

Figure 12: Scope for further improvement

7 Acknowledgements

We would like to thank Dr. Bernardis for her invaluable support and Dr. Paul for providing us an opportunity to work on this project.

References

[1] Madani, S., and Shapiro, J. (2000). Alopecia areata update. Journal of the American Academy of Dermatology, 42(4), 549-566.

[2] www.naaf.org

[3] Safavi, K. H., Muller, S. A., Suman, V. J., Moshell, A. N., and Melton III, L. J. (1995, July). Incidence of alopecia areata in Olmsted County, Minnesota, 1975 through 1989. In Mayo Clinic Proceedings (Vol. 70, No. 7, pp. 628-633). Elsevier.

[4] Bernardis, E., and Castelo-Soccio, L. (2018, January). Quantifying Alopecia Areata via Texture Analysis to Automate the SALT Score Computation. In Journal of Investigative Dermatology Symposium Proceedings (Vol. 19, No. 1, pp. S34-S40). Elsevier.