KASDI Merbah University - Ouargla Faculty of New Technologies of Information and Communication Department of Computer Science and Information Technology

Machine Learning 1: Blood vessel segmentation

, By: Moncef DJEZZAR, and Abdessamad SADOUDI
Under the supervision of: Dr. Belal KHALDI

Abstract: This study comparatively evaluates logistic regression and generalized linear models (GLMs) for the binary segmentation task of retinal blood vessel pixels from fundus imaging data. Accurate vessel segmentation is critical for diagnosing diabetic retinopathy and other eye diseases.

Keywords: Automated vessel segmentation, Diabetic retinopathy, Logistic regression, Generalized linear models (GLMs).

1. Introduction

The prevalence of diabetes and its ophthalmologic complication diabetic retinopathy underscores the pressing need for automated computational techniques to assist clinicians in disease screening and monitoring progression. A key step is the accurate segmentation of the retinal vasculature from fundus photographs, as morphological changes to the blood vessel network are indicative of disease status. While manual annotation by experts is time-consuming and subjective, automated vessel segmentation still poses challenges.

This study focuses on two established statistical modeling techniques: logistic regression and generalized linear models (GLMs). Logistic regression models the log-odds or logit of the binary output linearly based on the input features. GLMs extend this framework, allowing non-linear relationships between inputs and outputs via link functions, while modeling the outputs using exponential family distributions. For retinal vessel segmentation at the pixel level, these models take pixel intensity values from fundus images as inputs and output probabilities of each pixel being classified as vessel or non-vessel.

We conduct a rigorous comparative evaluation of logistic regression and GLMs on the public retina blood vessel segmentation dataset containing fundus images and corresponding manual vessel segmentation masks.

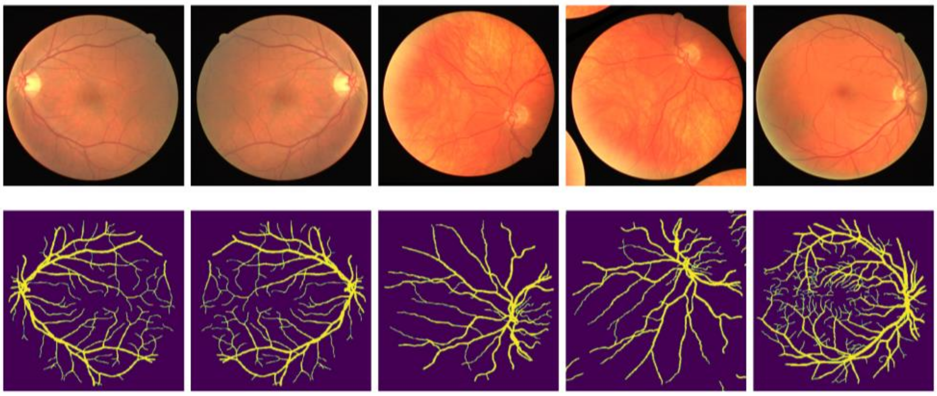
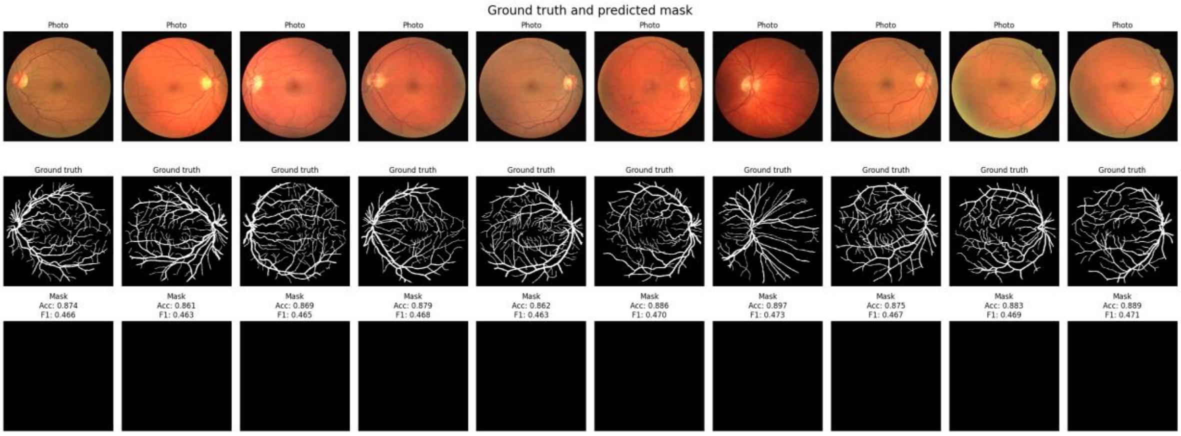


Figure: Retinal Fundus Images and Corresponding Masks

2. Naive Approaches in Model Training

This approach involves training the model directly on the available training dataset without any preprocessing or feature engineering.

2.1– Logistic Regression Model

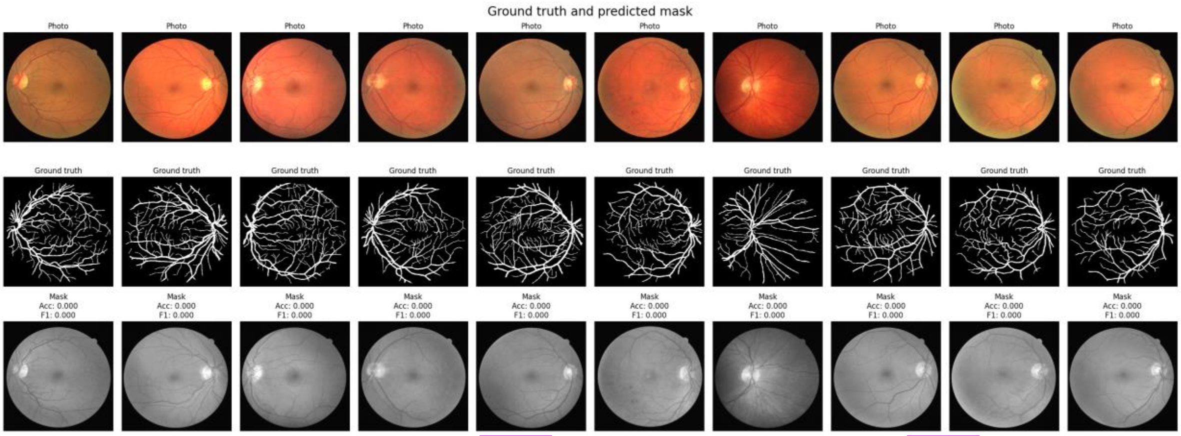


Global avg accuracy: 0.877 Global avg F1 score: 0.467

Figure: Performance of Logistic Regression Model on Raw Data

We see that the model does not work well when trained on raw, unprocessed data. The model makes a mistake by thinking that coloring everything black is the best solution, since the training data masks are mostly black.

2.2– Generalized Linear Model (GLM)



Global avg accuracy: 0.000 Global avg F1 score: 0.000

Figure: Performance of Generalized Linear Model (GLM) on Raw Data

Just like logistic regression, we can see that the model doesn't work very well when trained on raw, unprocessed data. The model's output looks very similar to the input, with almost no visible changes or improvements.

3. Improved Approach: Preprocessing, Feature Engineering, and Post-Processing

In contrast to the naive approaches discussed earlier, a more effective strategy involves a comprehensive pipeline that includes preprocessing, feature engineering, and post-processing steps.

3.1– Prepocessing

A. Color Channels

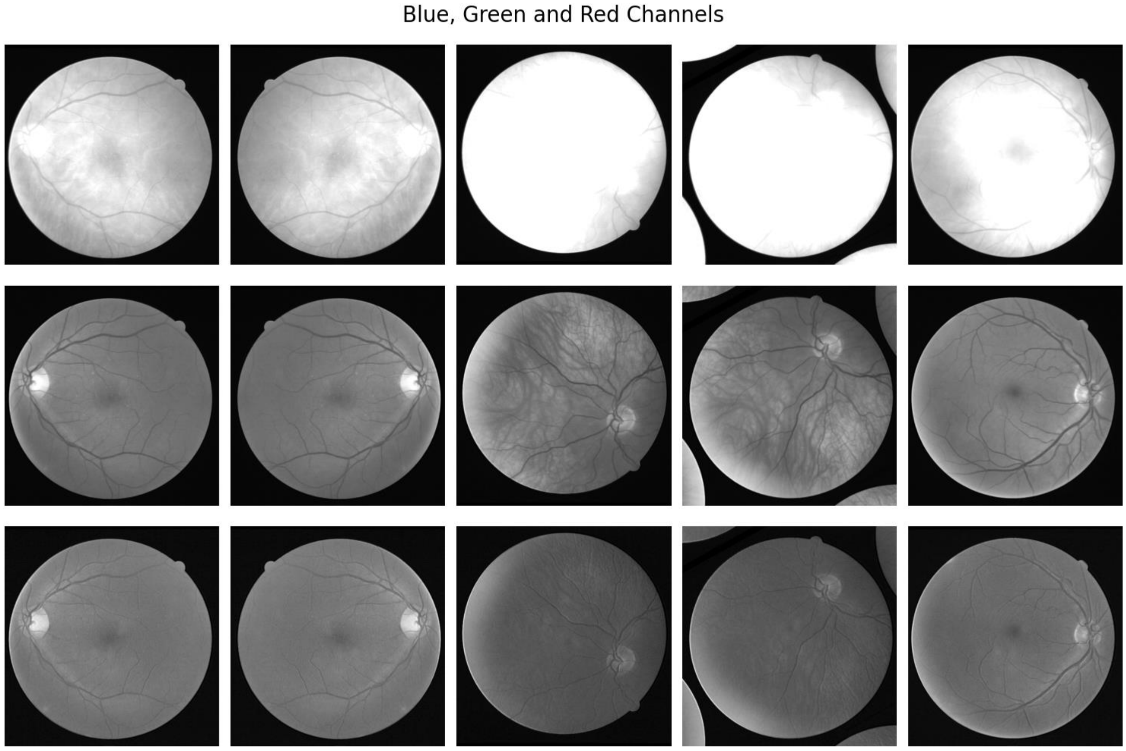


Figure: Color Channels in Fundus Images

We can see that the green color channel shows the blood vessels most clearly. Because of this, we will use the green channel instead of the original input image in the training process.

B. Contrast Enhancement

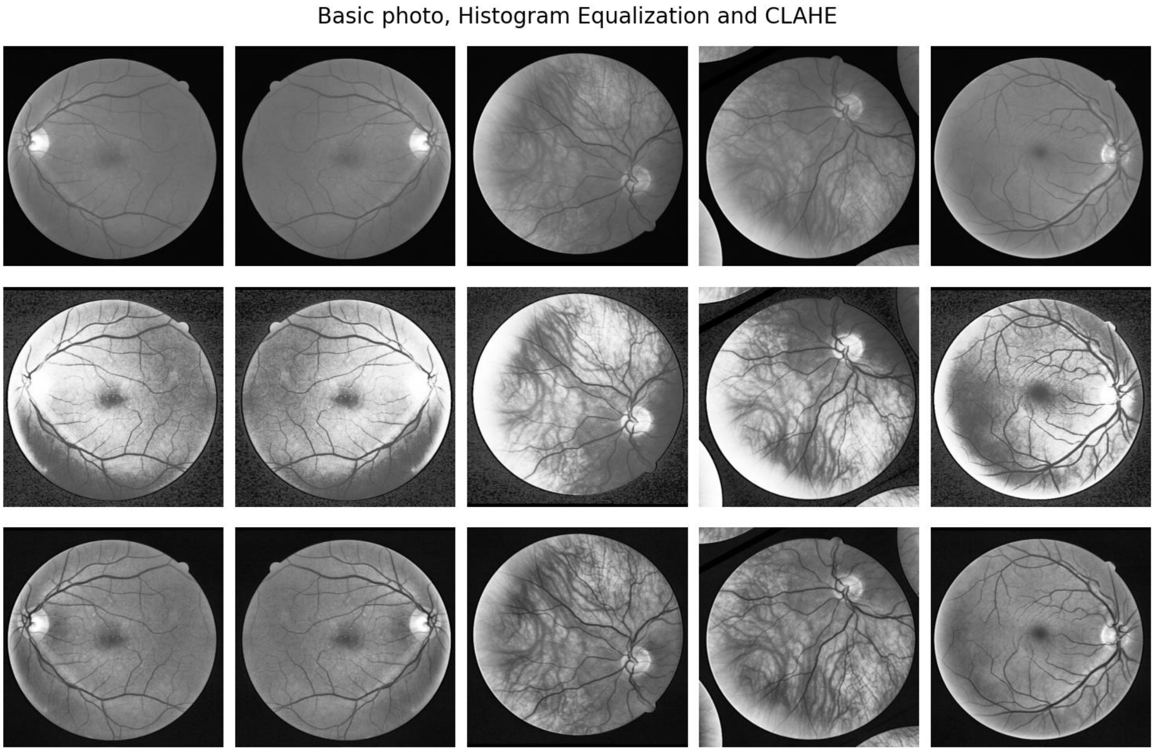


Figure: Contrast Enhancement Techniques

The techniques of histogram equalization and CLAHE (Contrast Limited Adaptive Histogram Equalization) made the blood vessels appear much clearer in the images. However, CLAHE seems to work even better than histogram equalization. We will use the green color channel from the original image, and further enhance its contrast using the CLAHE technique. This enhanced green channel image will be used as the training data.

3.2– Feature Engineering

In the feature engineering phase, we added a border around each image and its corresponding mask. The border size was half the intended patch size.

With these padded images and masks, we then extracted smaller 5x5 pixel patches from them in a non-overlapping manner, with a gap of 5 pixels between each patch.

For every extracted patch, we calculated two sets of features:

1. Color variance: Captured the variations in color within the patch.
2. Hu moments: Encoded information about the patch's shape and structure.

At the end of this process, we had two lists:

1. The first list contained the calculated color variance and Hu moment feature vectors for all the extracted patches from the training images.
2. The second list had the corresponding mask values for each of these patches.

Finally, we undersampled the lists to balance the dataset.

3.3– Models Training

A. Logistic Regression Model

Figure: Logistic Regression Output

The current model performs much better than the naive approach. However, we can see that there is a circle that significantly reduces the model's accuracy.

B. Generalized Linear Model (GLM)

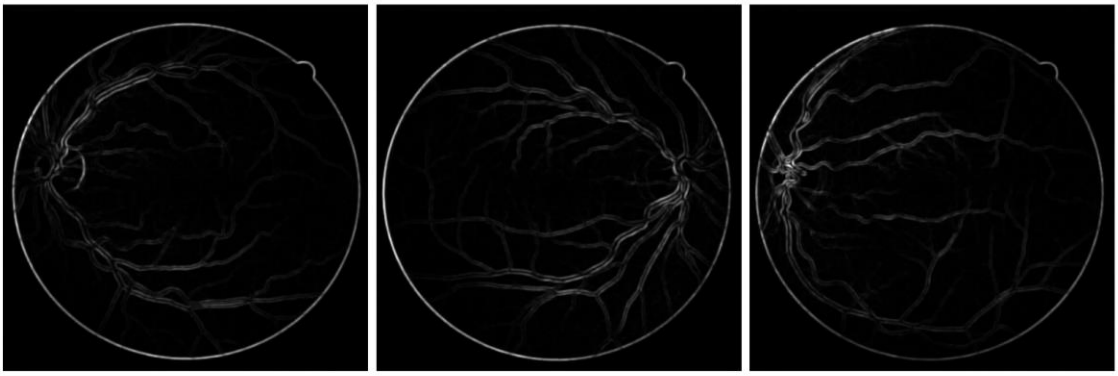


Figure: Generalized Linear Model (GLM) Output

Similar to logistic regression, the current model performs much better than the naive approach. However, just like logistic regression, we can see a circle that significantly reduces how accurate the model is. Additionally, the output image does not have enough contrast.

3.4– Post-Processing

A. Logistic Regression Model

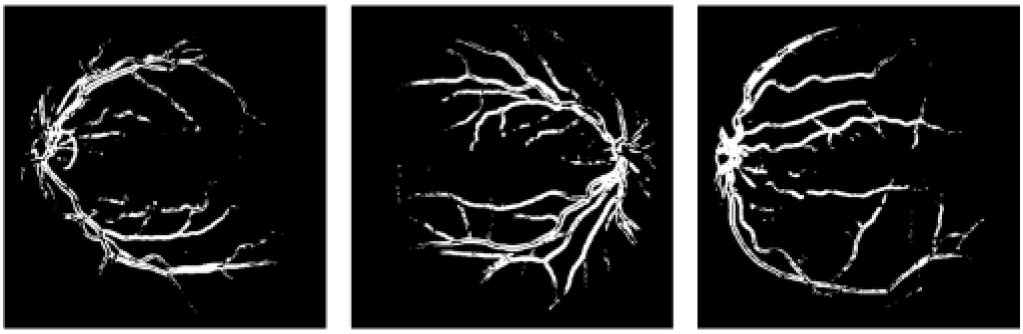


Figure: Logistic Regression Output (Circle Removed)

B. Generalized Linear Model (GLM)

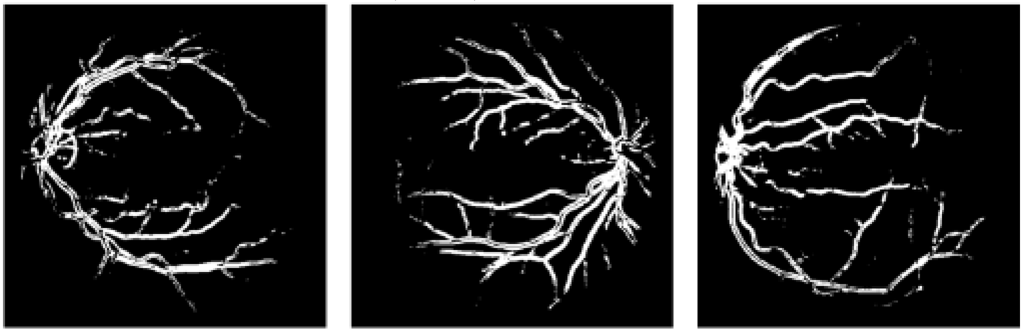
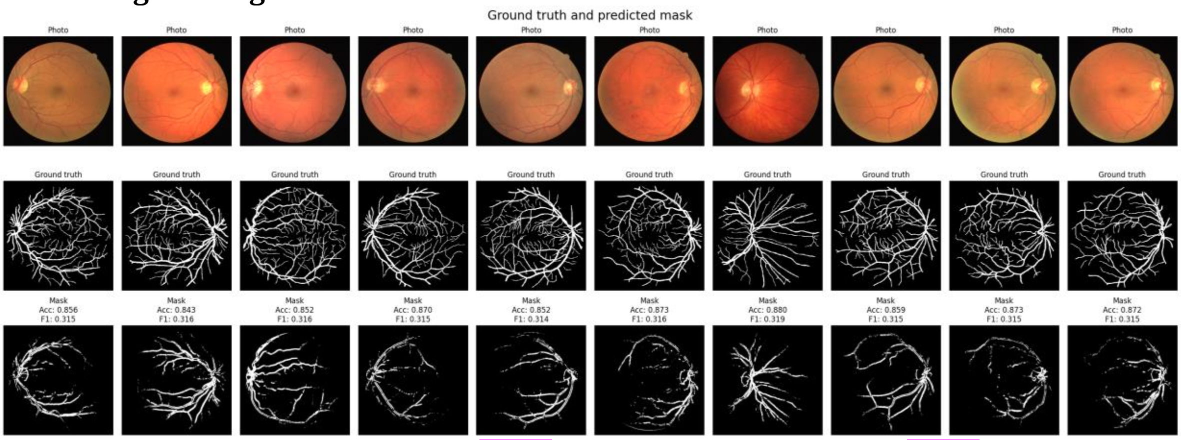


Figure: Generalized Linear Model (GLM) Output (Circle Removed, Enhanced Contrast)

3.5– Evaluation

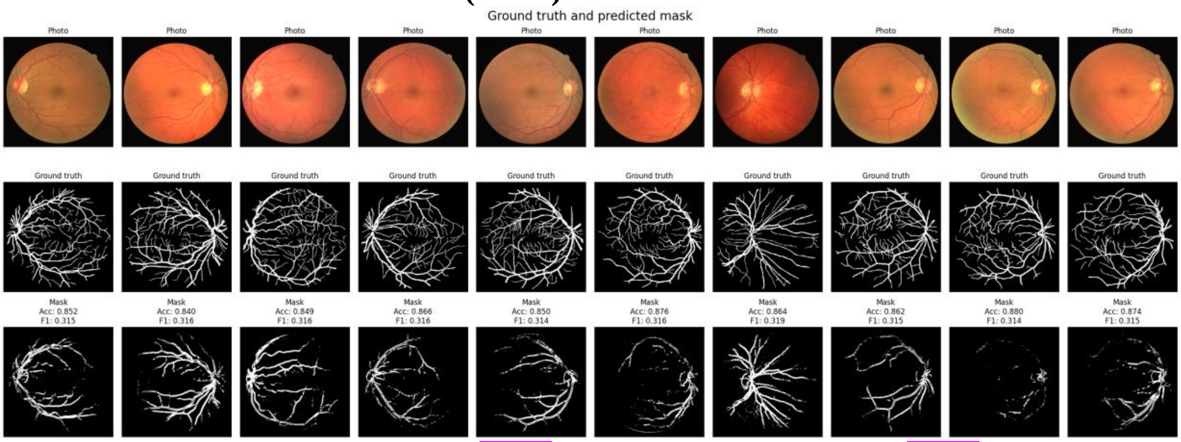
A. Logistic Regression Model



Global avg accuracy: 0.865 Global avg F1 score: 0.315

Figure: Performance of Logistic Regression Model (Enhanced Approach)

B. Generalized Linear Model (GLM)



Global avg accuracy: 0.864 Global avg F1 score: 0.315

Figure: Performance of Generalized Linear Model (GLM) (Enhanced Approach)

4. Comparative Analysis

4.1– Logistic Regression Model

A. Strengths:

- Logistic regression is a well-established and interpretable machine learning algorithm that is suitable for binary classification tasks like image segmentation.
- The model provides probabilities for each pixel, which can be thresholded to obtain the final binary segmentation. The threshold can be adjusted based on the desired output.

B. Weaknesses:

- Logistic regression is a linear model, which may not be able to capture complex non-linear relationships in the data as effectively as more flexible models like neural networks.
- The global average accuracy (0.865) and F1-score (0.315) suggest that the model may not be performing as well as desired, potentially due to the complexity of the segmentation task.

4.2– Generalized Linear Model (GLM)

A. Strengths:

- The GLM model may be able to capture non-linear relationships in the data better than the logistic regression model, as it does not have the same linearity constraint.
- The global average accuracy (0.864) and F1-score (0.315) are comparable to the logistic regression model.

B. Weaknesses:

- The GLM model may be more computationally expensive than the logistic regression model, especially when working with large-scale datasets or high-resolution images.
- The performance of the GLM model may be more sensitive to the choice of the power parameter, which determines the distribution family, and the link function, which determines the relationship between the predictor and the response.

5. Conclusion

The two models produce relatively comparable segmentation results, with a slight difference in the details of the segmented regions.