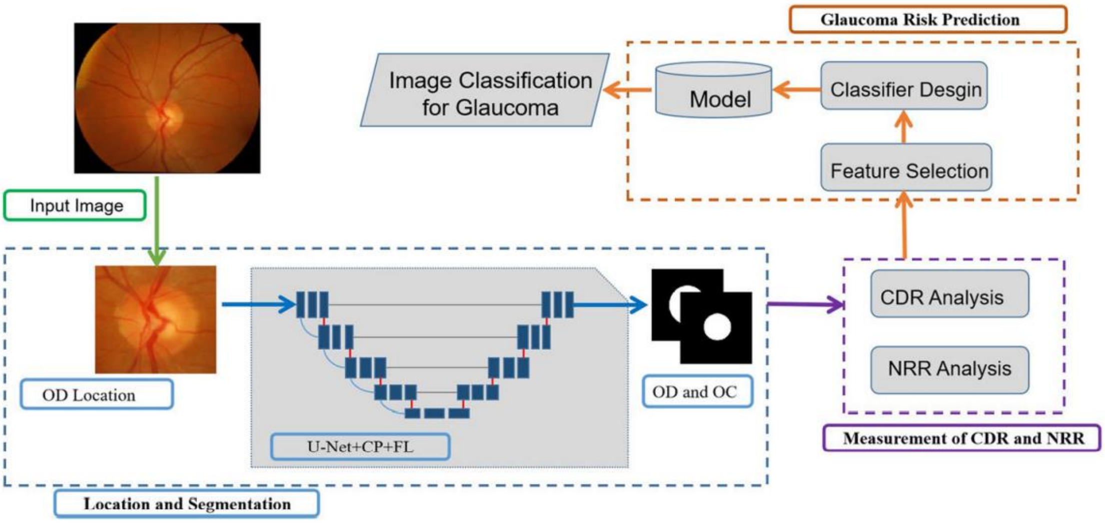
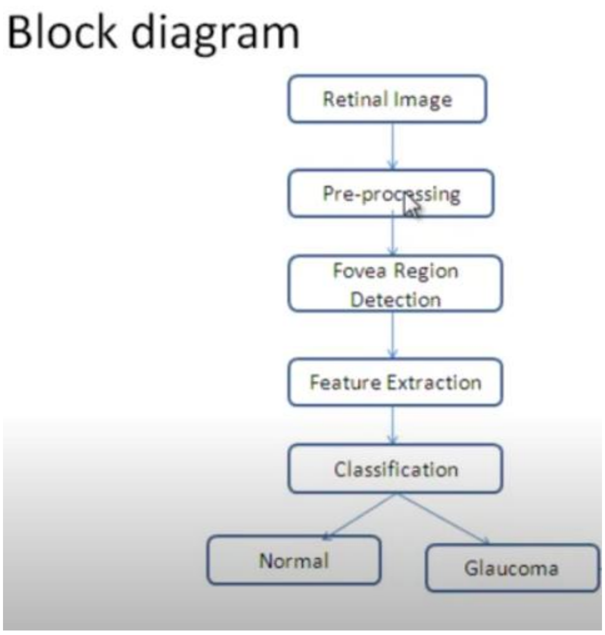
Architectural Decisions Document

PREDICTING OCCURRENCE OF GLAUCOMA FROM RETINA FUNDUS IMAGES

1 Architectural Components Overview

- Anaconda Navigator, Jupyter notebook 5.7.8, Python 3, Anaconda Cloud




Block diagram

ARCHITECTURAL CHOICES

- Anaconda Navigator, Jupyter notebook 5.7.8, Python 3
- Anaconda Cloud
- Image loading / transformation:
  - OS module for interacting with operating system
  - TKinter for standard Graphical User Interface (GUI) package (filedialog)
  - Scikit-image for importing data and Scikit-image.feature 'greycomatrix' and 'greycoprops' for extracting features from images
  - SciPy signal for Gaussian window filter
  - OpenCV (cv2.imread) for reading images and (cv2.resize) for resizing images and (cv2.split) for splitting RGB into 3 channels (Red, Green, Blue) and for Morphological segmentation.
- Traditional Machine learning algorithm: Scikit-learn for SVM and Random Forest
- Deep learning algorithm: tensorflow with tensorflow.keras for CNN Classification

1.1 Data Source

1.1.1 Technology Choice

The data was downloaded from Kaggle (https://www.kaggle.com/sshikamaru/glaucoma-detection)

1.1.2 Justification

Primary reason to download from Kaggle was availability and ease of use.

1.2 Enterprise Data

1.2.1 Technology Choice

ORIGA-light : An Online Retinal Fundus Image Database for Glaucoma Analysis and Research

1.2.2 Justification

This Database consists of 650 images (including 168 glaucomatous images and 482 randomly selected nonglaucoma images) from SiMES study. Each image is segmented and annotated by trained professionals from Singapore Eye Research Institute. Research done by 8 authors: Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, Tien Yin Wong

1.3 Streaming analytics

1.3.1 Technology Choice

NA

1.3.2 Justification

NA

1.4 Data Integration

1.4.1 Technology Choice

Not used

1.4.2 Justification

Not used

1.5 Data Repository

1.5.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case

this component is not needed justify below.

1.5.2 Justification

Please justify your technology choices here.

1.6 Discovery and Exploration

1.6.1 Technology Choice

The following Python 3 libraries were used for Data Exploration and Visualization: -

- Image loading / transformation:
    - OS module for interacting with operating system
    - TKinter for standard Graphical User Interface (GUI) package (filedialog)
    - Scikit-image for importing data and Scikit-image.feature ‘greycomatrix’ and ‘greycoprops’ for extracting features from images

- SciPy signal for Gaussian window filter
- OpenCV (cv2.imread) for reading images and (cv2.resize) for resizing images and (cv2.split) for splitting RGB into 3 channels (Red, Green, Blue) and for Morphological segmentation.
- Traditional Machine learning algorithm: Scikit-learn for SVM and Random Forest
- Deep learning algorithm: tensorflow with tensorflow.keras for CNN Classification

1.6.2 Justification

The size of the dataset was the key factor in deciding data exploration tools.

The current data small enough to be processed on a single computer ruling out the need for

distributed processing (Spark, pyspark)

1.7 Actionable Insights

1.7.1 Technology Choice

The following Python 3 libraries were used for Data Exploration and Visualization: -

Pandas,

Keras,

Tensoflow.

1.7.2 Justification

CNN (convolution neuro network) based algorithm was used as a reference for the Tree based model. Easiest

and Fastest implementation is possible in keras. Tensorflow is the backend.

Best results with deep learning model - Convolutional Neural Network (CNN) in Keras, TensorFlow.

- Batch size: 64
- Number of Epochs: 20
- Architecture: Convolution layers followed by dense layers
- Optimizer: Adam
- Test loss: sparse categorical crossentropy 2.0893890133e-05
- Accuracy on test data: 93.83%

I tried 2 traditional machine learning models:

- SVM (support vector machine), accuracy on test data: 71.53%
- Random Forest, accuracy on test data: 93.50%

1.8 Applications / Data Products

1.8.1 Technology Choice

A Jupyter notebook based report was generated

1.8.2 Justification

As only the correlating factors needed to be identified Jupyter notebook based report was

consider sufficient.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

None

1.9.2 Justification

NA