Diabetic Retinopathy Classification and Interpretation using Deep Learning Techniques

Jordi de la Torre

Doctorate Program of Computer Science and Mathematics of Security
Supervisors: Dra. Aïda Valls and Dr. Domènec Puig
Universitat Rovira i Virgili

March 12, 2019

Outline

Part I

Introduction and Background

Outline

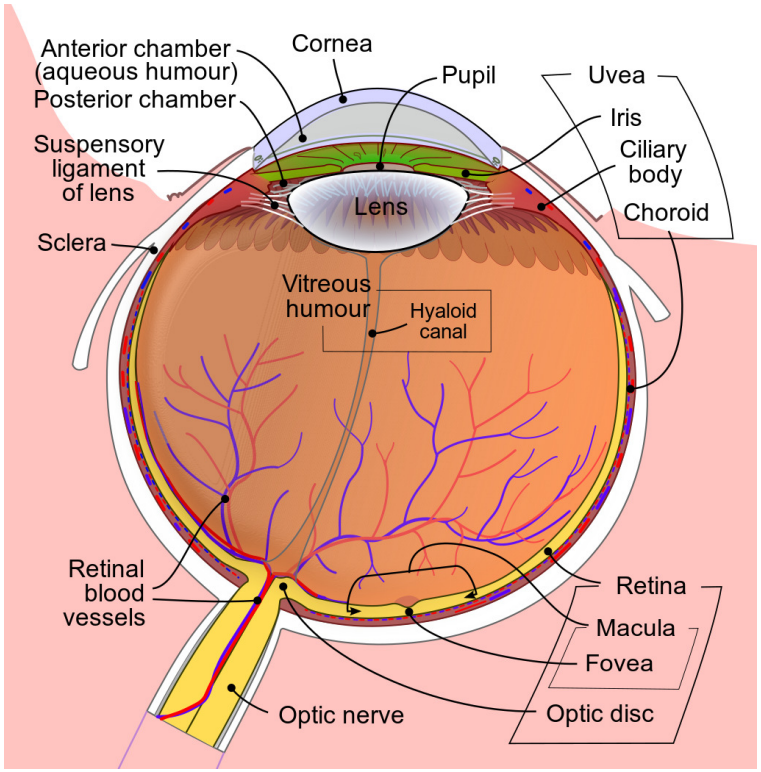
Thesis Objective

Design a self-explainable method for automatic diabetic retinopathy disease grading, based on the analysis of retina fundus images with an accuracy close to the human experts in the field.

Diabetic Retinopathy

Eye Structure

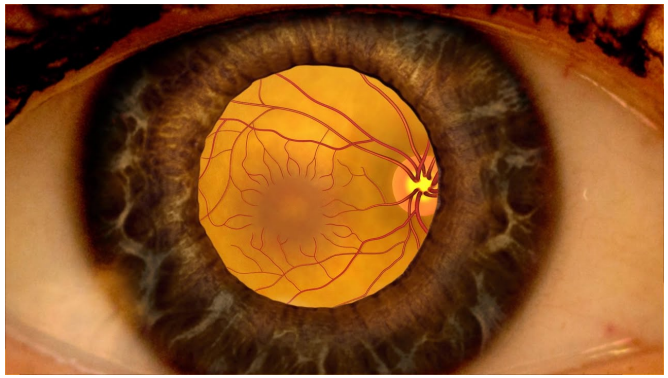
- Light passes through the cornea, iris and lens reaching the internal structures of the eye.
- It impacts the back of the eye, where retina is located.
- Light activates a set of sensory elements in retina.
- The signal of these sensory elements is transported through the optic nerve to the brain.



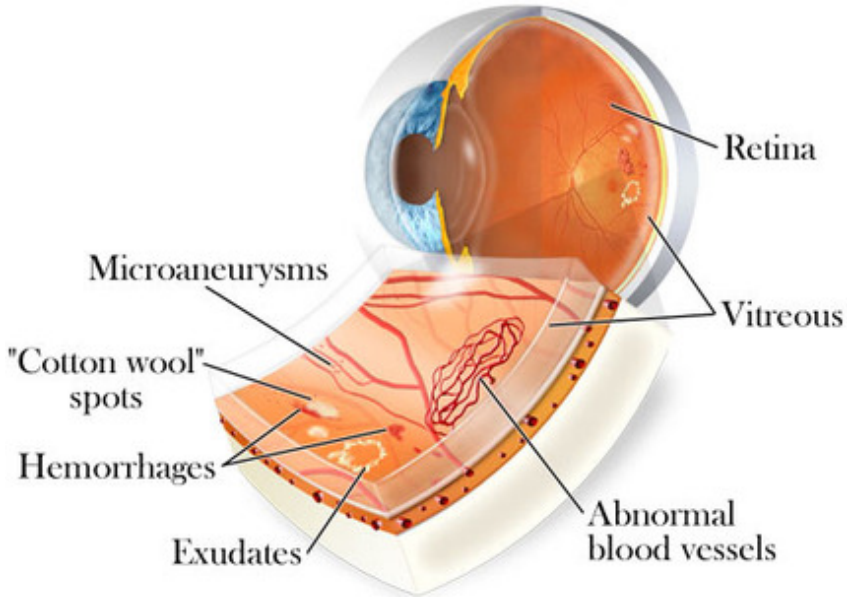
Diabetic Retinopathy

What diabetic retinopathy is?

- Disease derived from diabetic condition.
- Affects to the sensory elements of the retina.
- Produced by the deterioration of retina irrigation capilars.
- If not treated, it can cause vision quality loss and eventually blindness.
- Typical lesions: microaneurysms, "cotton wool" spots, abnormal neovascularization, hemorrhages.



Diabetic Retinopathy



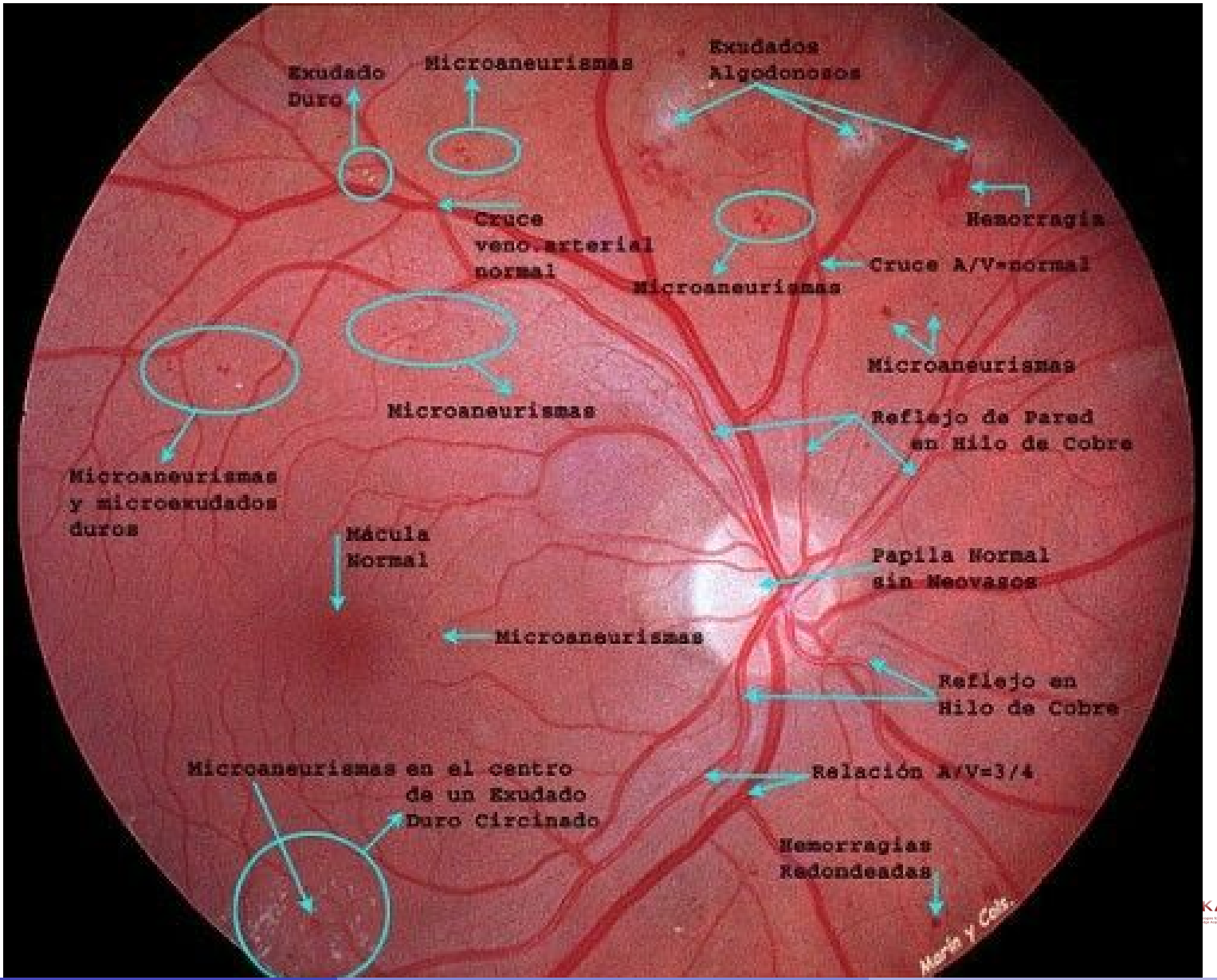
Diabetic Retinopathy

Evaluation Technique: Fundus photography

It allows the visualization of main structures present in the back of eye interior.

Diabetic Retinopathy

Typical DR lesions



Diabetic Retinopathy Grade Classification

Grading table

INTERNATIONAL CLINICAL DIABETIC RETINOPATHY DISEASE SEVERITY SCALE

Proposed Disease Severity Level	Findings Observable upon Dilated Ophthalmoscopy
No Apparent Retinopathy	€ No abnormalities
Mild Non-Proliferative Diabetic Retinopathy	Microaneurysms only
Moderate Non-Proliferative Diabetic Retinopathy	More than just microaneurysms but less than Severe NPDR
Severe Non-Proliferative Diabetic Retinopathy	Any of the following: € More than 20 intraretinal hemorrhages in each of 4 quadrants € Definite venous beading in 2+ quadrants € Prominent IRMA in 1+ quadrant And no signs of proliferative retinopathy
Proliferative Diabetic Retinopathy	One or more of the following: € Neovascularization € Vitreous/preretinal hemorrhage

RiVI

Evaluation Measures

Binary Classification

Evaluation Methods used for binary classification when a "true gold standard" is available:

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= \frac{1}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Source: `https://en.wikipedia.org/wiki/Sensitivity_and_specificity`

Evaluation Measures

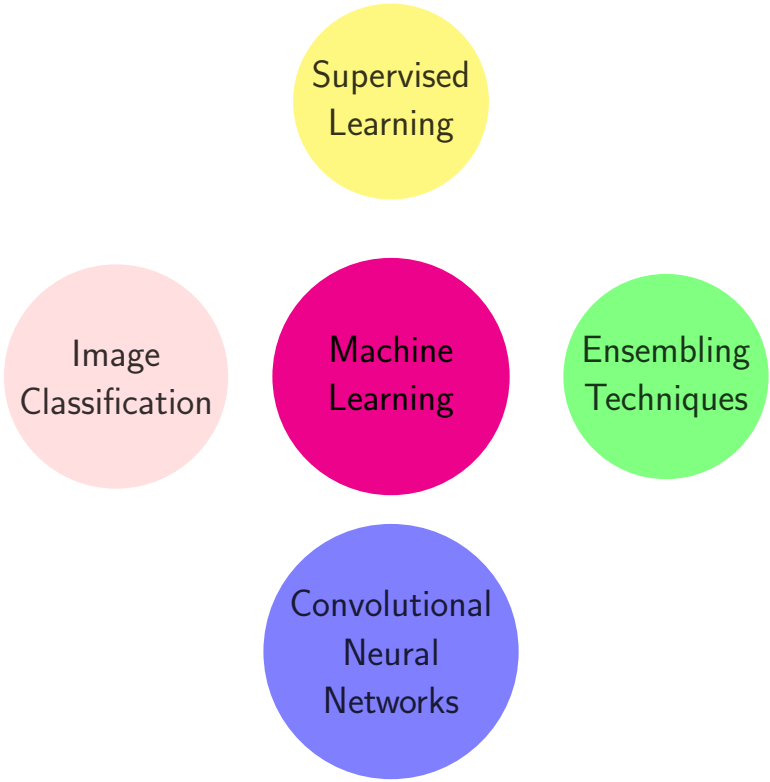
Inter-rater agreement

- Used for measuring agreement between raters
- Kappa (binary classification, no gold standard) $\kappa = \frac{P_0 - P_e}{1 - P_e}$
- Weighted Kappa (ordinal categories, penalization grows with distance) $\kappa_w = \frac{\sum_{i,j} \omega_{i,j} O_{i,j}}{\sum_{i,j} \omega_{i,j} E_{i,j}}$
- Intra-class correlation (when outcome measured on a continous scale)

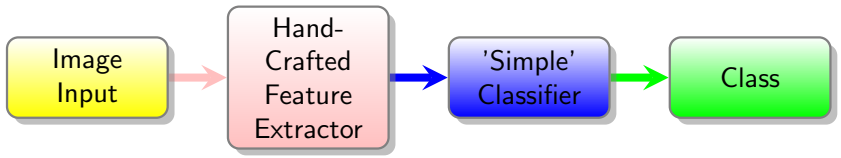
Table: Table for interpretation of Weighted Kappa, after Landis & Koch (1977)

κ	Strength of agreement
≤ 0.20	Poor
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Good
$0.81 - 1.00$	Very good

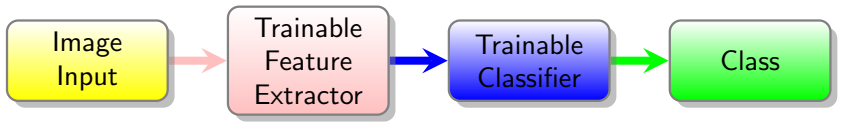
Methods Used in this thesis



Traditional pattern recognition scheme:



Deep Learning pattern recognition scheme (end-to-end learning):

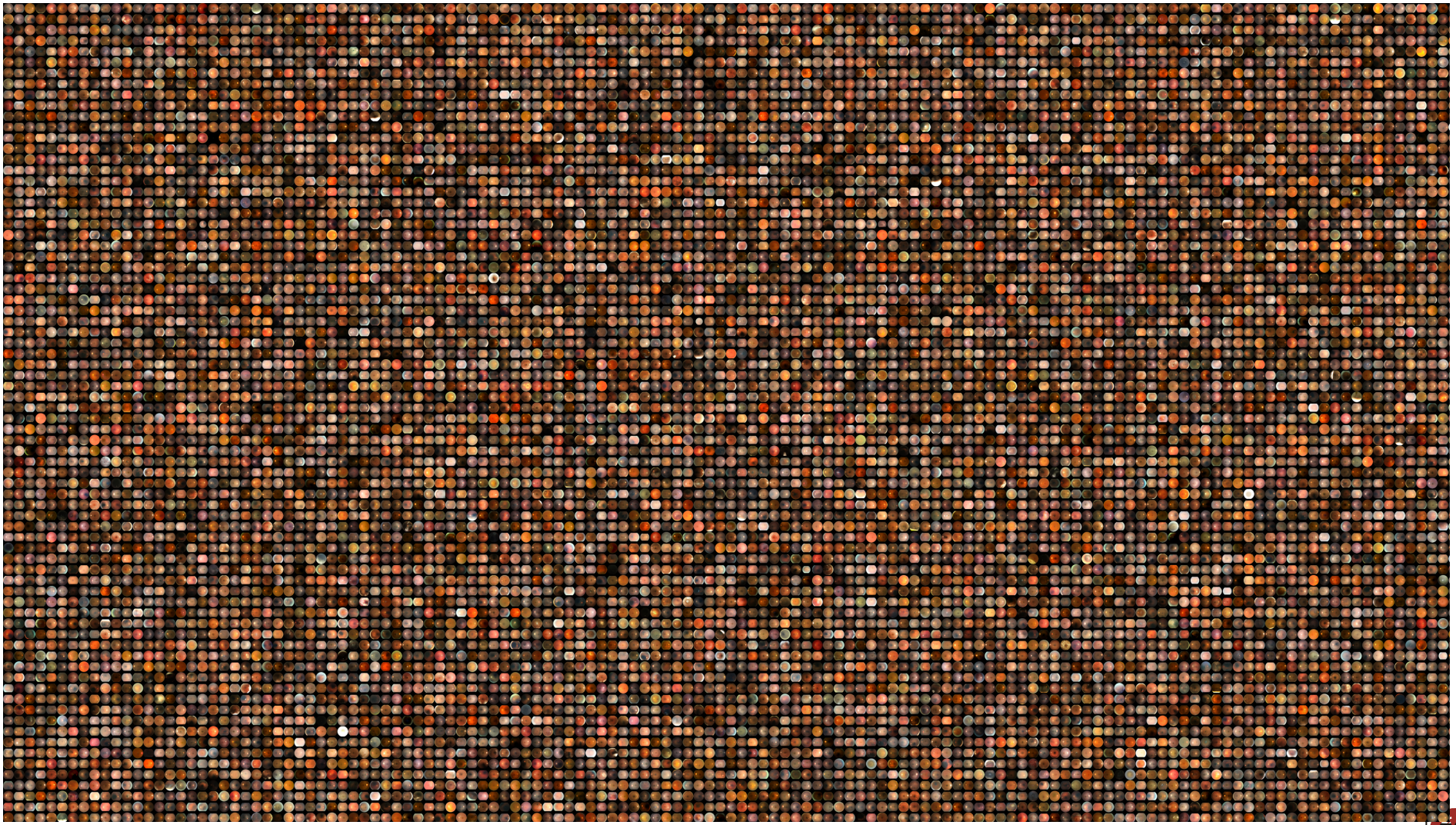


The dataset: EyePACS

Whole dataset of near 88,692 images, 10x10

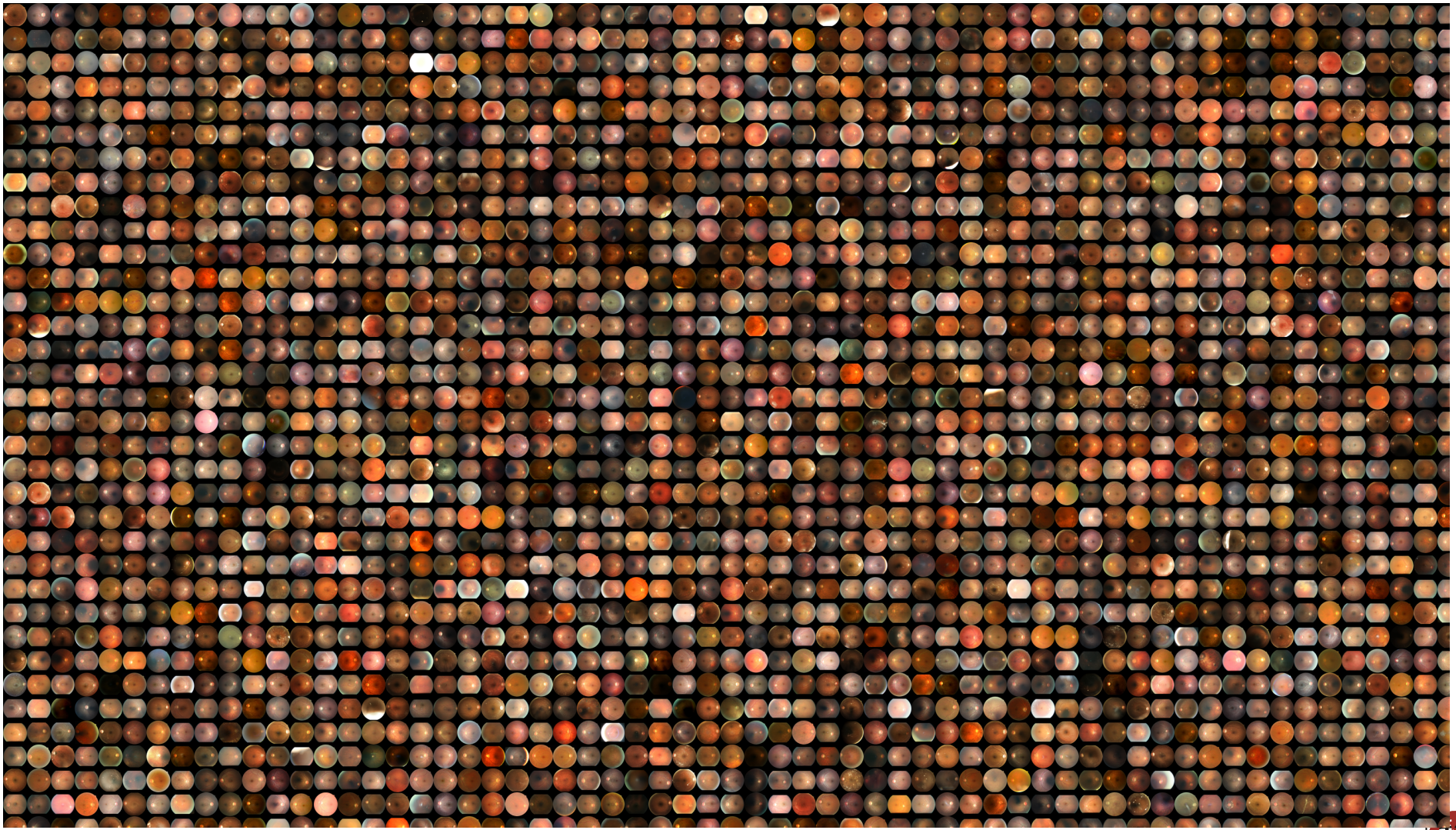
The dataset: EyePACS

10,000 random images, 25x25



The dataset: EyePACS

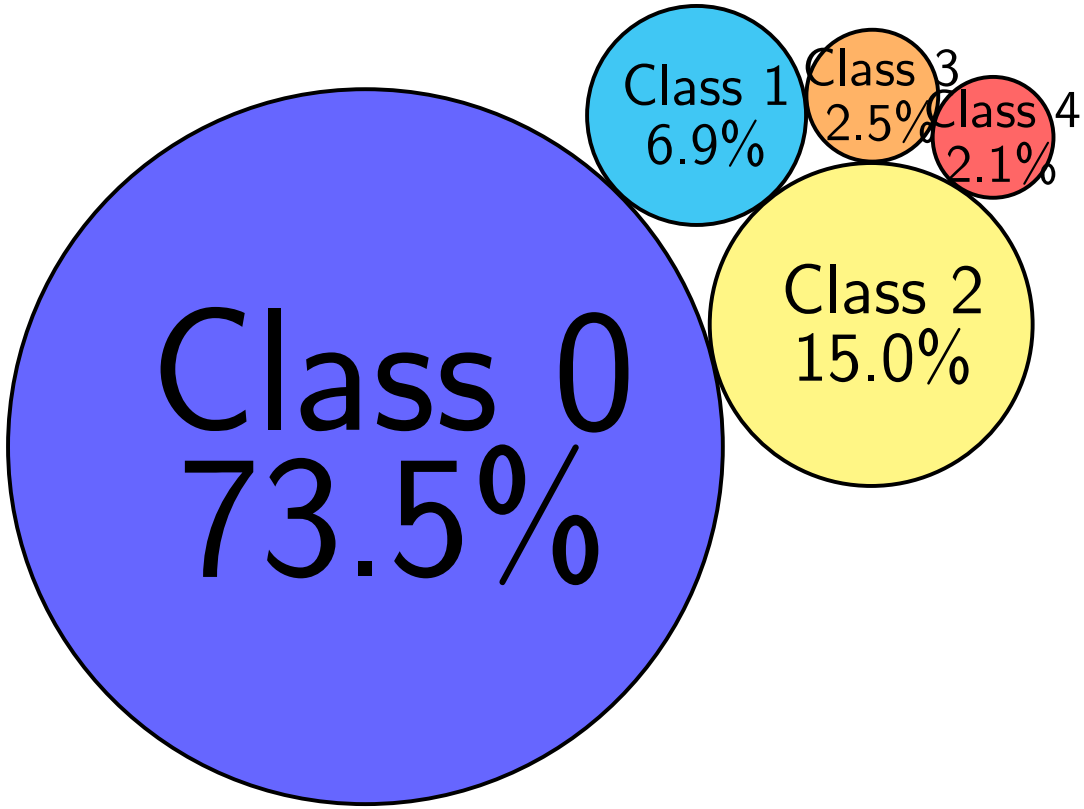
2,000 random images, 50x50



The dataset: EyePACS

500 random images, 100x100

EyePACS dataset



- 88,692 retina fundus images of differing sizes, illumination conditions, quality.
- 44,346 different patients. For every patient: right and left eye images available.

The dataset

Minimal data pre-processing

Minimal data preprocessing applied (optimization of hardware resources, deep learning optimization methods are computer intensive tasks):

- Remove borders
- Standardization of the image size (downsizing: 128, 256, 512, etc.) depending on the model requirements.

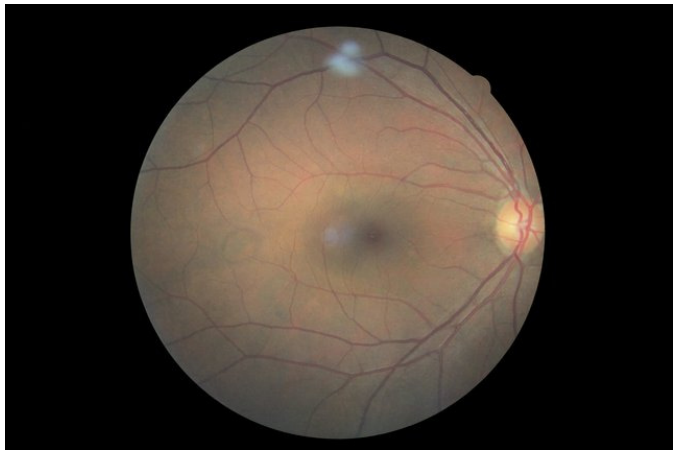


Figure: Original 4752x3168 pixels

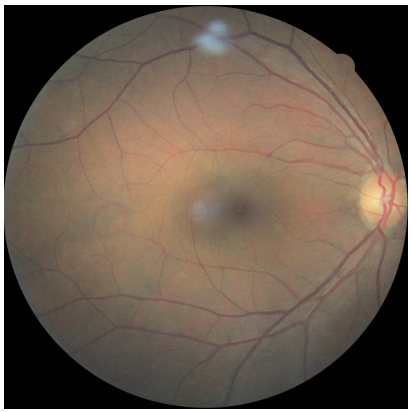


Figure: Preprocessed: trimmed + resized

Part II

Classification

Outline

Classification - Preliminary models

Mathematical formalization of the classification problem

- Find a function f: $\mathbb{R}^{CxHxW} \mapsto \mathbb{R}^n$ that maximizes a objective function (where $CxHxW \gg n$). Optimization.
- Images are high dimensional objects with highly correlated local points. Function proven to exploit these characteristics: a deep convolutional neural network.
- Objective function in neural network argot called cost function.
- Standardized cost function for classification: logarithmic loss (log-loss).
- Evaluation metric: quadratic weighed kappa (QWK).

Classification - Preliminary models

Preliminary Models - Key points of this work

- Data augmentation techniques were necessary to balance the classes and to increment the generalization capabilities of the model (brightness,contrast, rotations).
- Due to hardware limitations only a part of the input was feeded to the network (about 71% of the useful information), requiring various evaluations and ensembling on test time to increase performance.
- Probabilistic combination of results of both eyes (Bayes rule) helps improve further performance. (Thesis contribution)

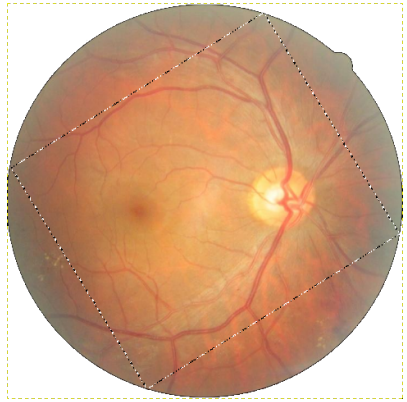
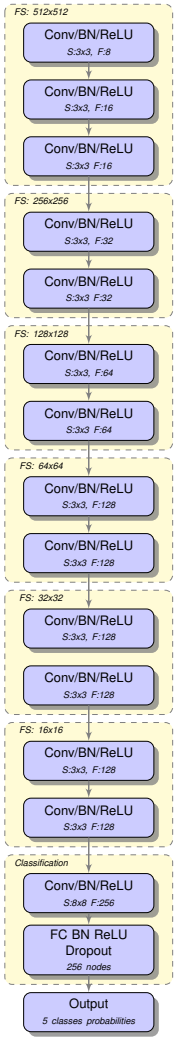


Figure: Input data selection

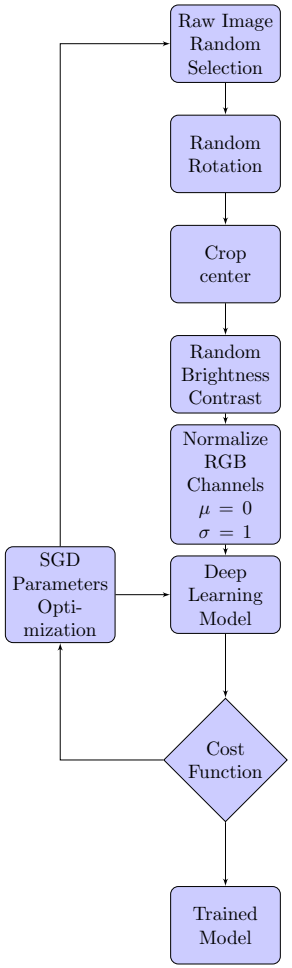
Classification - Preliminary models

Preliminary models - Model and training overview (CCIA 2016)

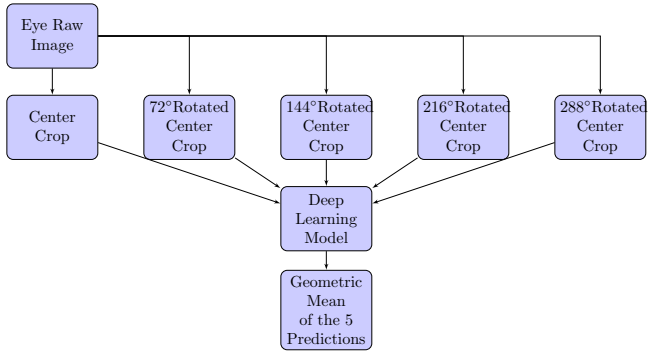
Model



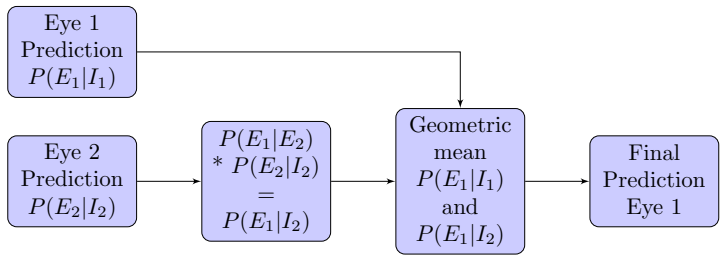
Training



Evaluation



Combination



Classification - Preliminary models

Preliminary models - Results

Layers	Input size	$\kappa_{test,alone}$	$\kappa_{test,combined}$
12	(3,128,128)	0.488	0.577
14	(3,256,256)	0.636	0.660
16	(3,384,384)	0.668	0.730
16	(3,512,512)	0.725	0.769

Table: Comparison of results obtained with and without probabilistic combination

Classification - Preliminary models

Summary of the methods used

- Data augmentation techniques
- Stochastic Gradient Descent
- Logarithmic loss function
- Partial input information due to hardware limitations
- Ensembling techniques with different versions of the same image (geometric mean of 5 different evaluations)
- Ensembling techniques for combination of the information of both eyes (Bayes rule)

Outline

QWK: A new loss function for ordinal regression

Mathematical summary of the paper (Pattern Recognition Letters, 2017)

Model function: $p = f(I) \quad where \quad I \in \mathbb{R}^{CxHxW}, \quad p \in \mathbb{R}^n$

Log-loss optimization:

$$min \quad C = \sum_{i=1}^{BS} \sum_{j=1}^{n} t_{i,j} log(p_{i,j})$$

First order derivative:

$$\frac{\partial C}{\partial p_{i,j}} = \frac{t_{i,j}}{p_{i,j}}$$

QWK-loss optimization:

$$max \quad \kappa = \frac{\sum_{i,j} \omega_{i,j} O_{i,j}}{\sum_{i,j} \omega_{i,j} E_{i,j}}$$

$$min \quad C = log(1-\kappa), \quad \omega_{i,j} = \frac{(i-j)^2}{(n-1)^2}$$

First order derivative: (next slide)

QWK: A new loss function for ordinal regression

Mathematical summary of the paper (Pattern Recognition Letters, 2017)

QWK first order derivative:

$$\frac{\partial \mathcal{L}}{\partial y_m} = \frac{1}{\mathcal{N}} \frac{\partial \mathcal{N}}{\partial y_m} - \frac{1}{\mathcal{D}} \frac{\partial \mathcal{D}}{\partial y_m}$$

where:

$$\frac{\partial \mathcal{N}}{\partial y_m} = \begin{pmatrix} \omega_{t_1,1} & \omega_{t_1,2} & \dots & \dots & \omega_{t_1,C} \\ \omega_{t_2,1} & \omega_{t_2,2} & \dots & \dots & \omega_{t_2,C} \\ \dots & \dots & \dots & \dots & \dots \\ \omega_{t_N,1} & \omega_{t_N,2} & \dots & \dots & \omega_{t_N,C} \end{pmatrix}$$

$$\frac{\partial \mathcal{N}}{\partial y_m(X_k)} = \omega_{t_k m}$$

$$\frac{\partial \mathcal{D}}{\partial y_m(X_k)} = \sum_{i=1}^{C} \hat{N}_i \omega_{i,m}$$

$$\frac{\partial \mathcal{D}}{\partial y_m} = \begin{pmatrix} \sum_{i=1}^{C} \hat{N}_i \omega_{1,i} & \dots & \dots & \dots & \sum_{i=1}^{C} \hat{N}_i \omega_{C,i} \\ \sum_{i=1}^{C} \hat{N}_i \omega_{1,i} & \dots & \dots & \dots & \sum_{i=1}^{C} \hat{N}_i \omega_{C,i} \\ \dots & \vdots & \dots & \dots & \dots \\ \sum_{i=1}^{C} \hat{N}_i \omega_{1,i} & \dots & \dots & \dots & \sum_{i=1}^{C} \hat{N}_i \omega_{C,i} \end{pmatrix}$$

$$m \in \{1, 2, ..., C\}$$

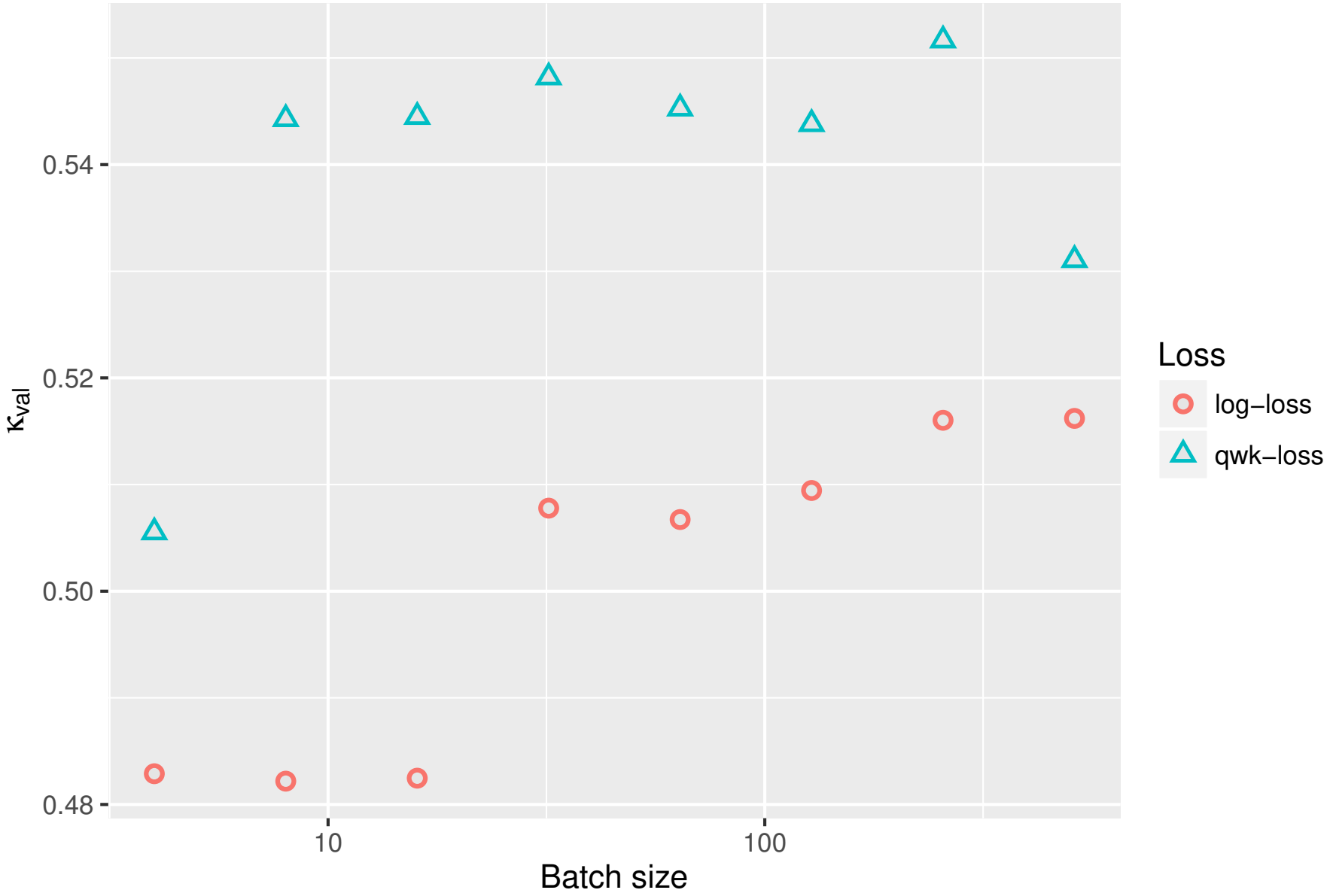
QWK: A new loss function for ordinal regression

Pattern Recognition Letters, 2017

- Three different multi-class classification problems using as evaluation metric QWK were trained using QWK-loss and log-loss.
- Different neural networks were tested: a linear classifier, a shallow neural network of 2-3 layers and a deep neural network of up to 16 layers.
- Optimizing QWK reported in all the models and problems increases of performance in the test set of about 5 to 10% over the conventional training method.

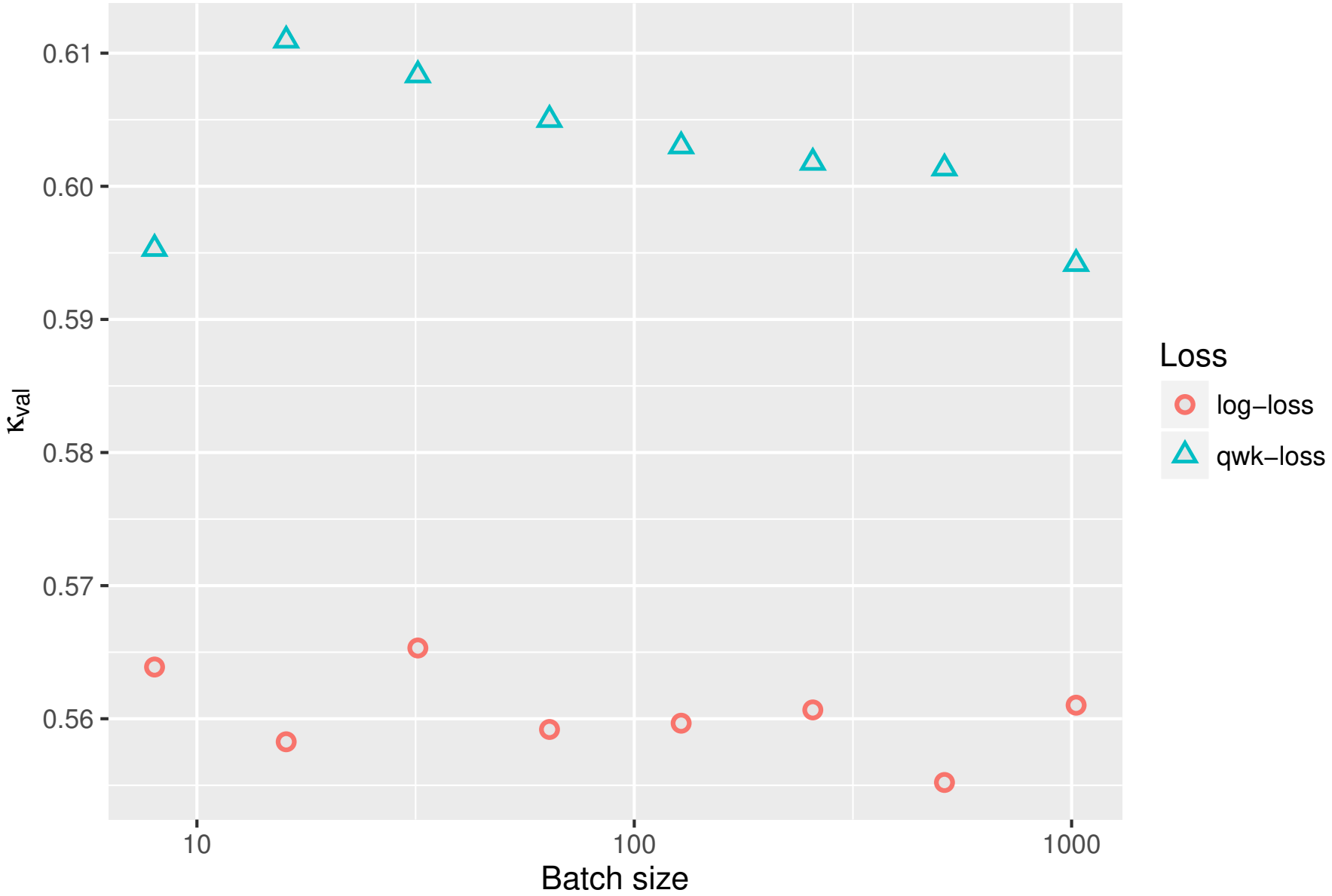
QWK: A new loss function for ordinal regression

Results for "Search Results Relevance" Case Study



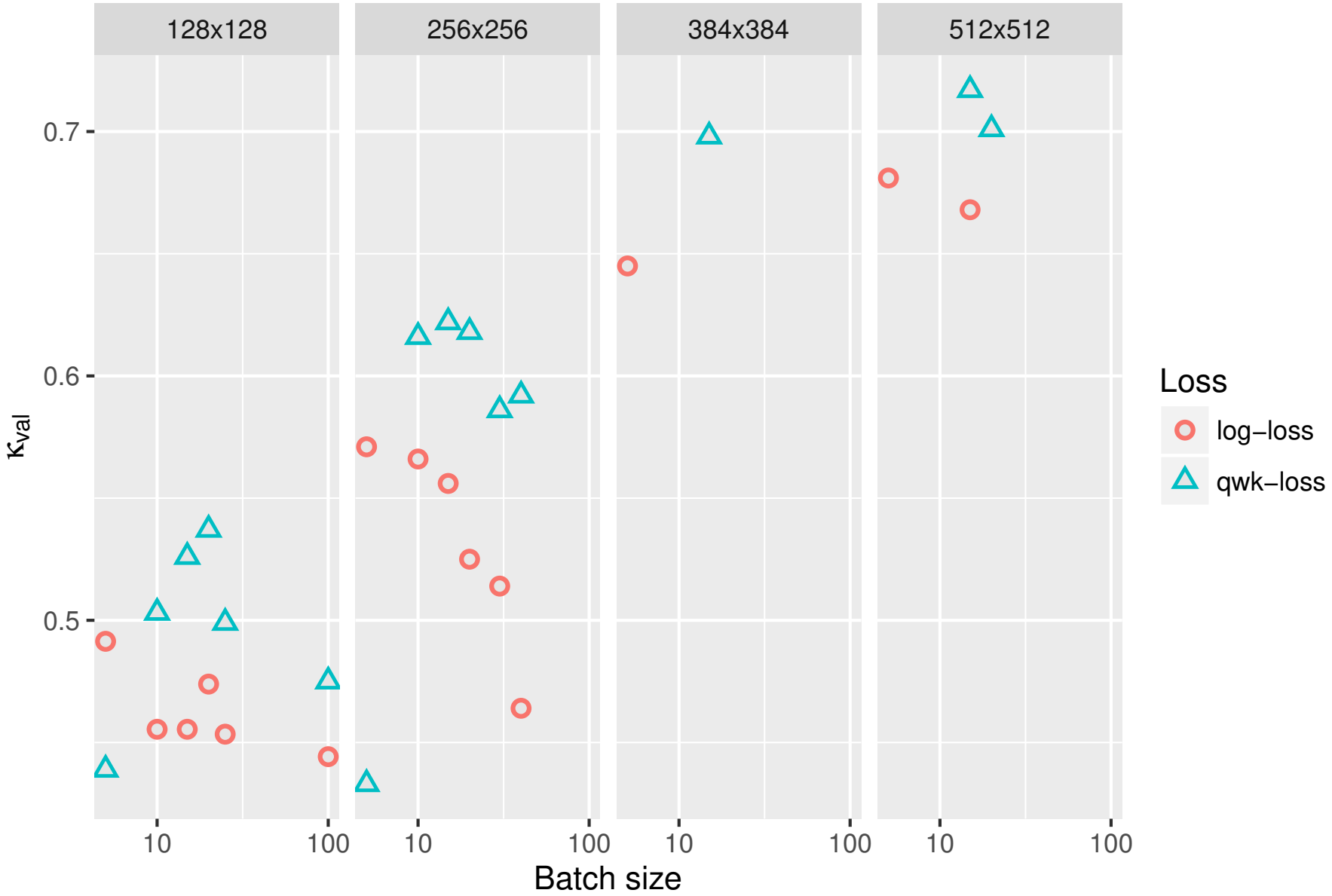
QWK: A new loss function for ordinal regression

Results for "SNN Insurance Assessment" Case Study



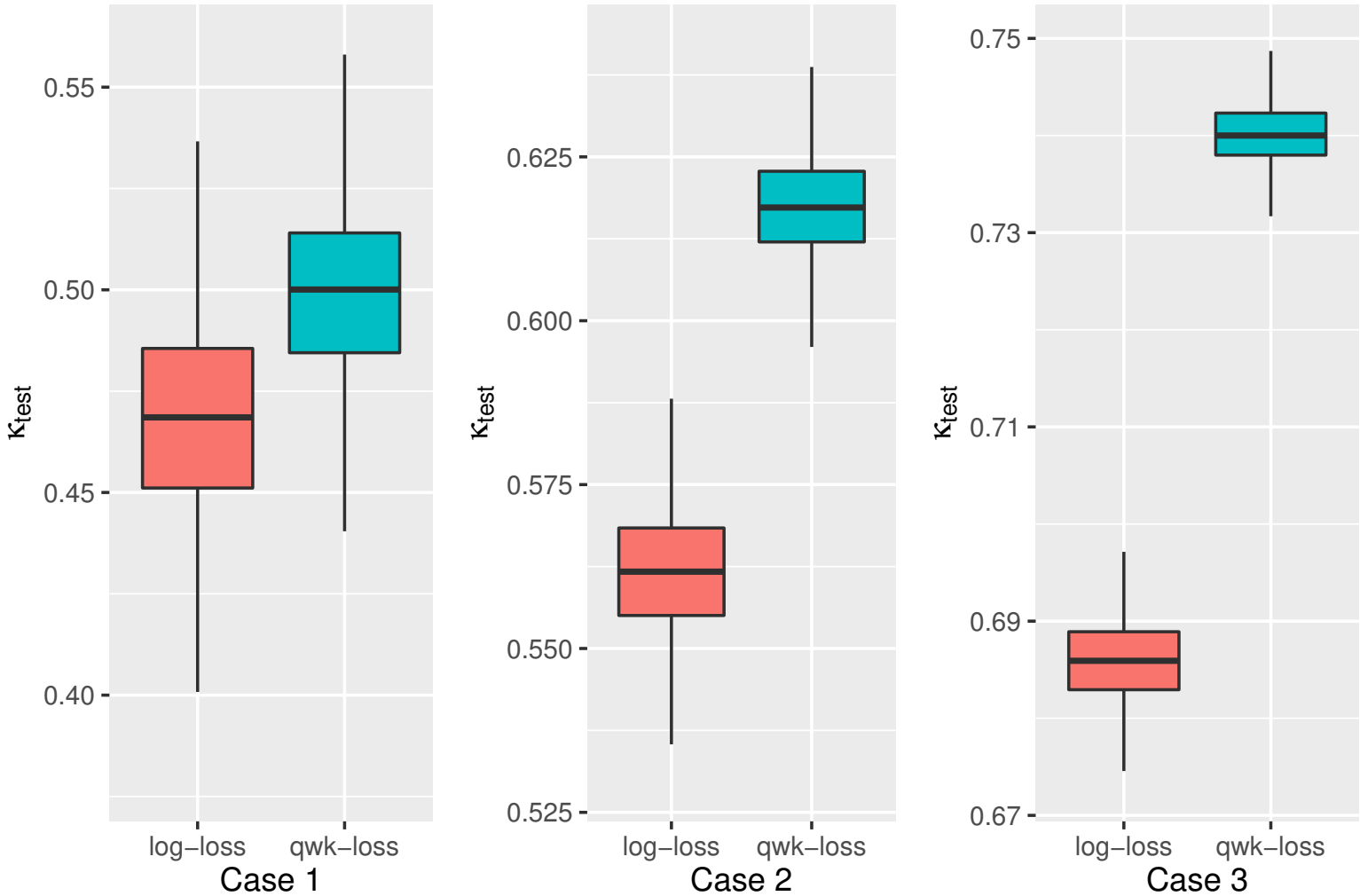
QWK: A new loss function for ordinal regression

Results for "Diabetic Retinopathy Disease Grading" Case Study



QWK: A new loss function for ordinal regression

Test set confidence intervals for each case study



Outline

Classification - Enhanced Models

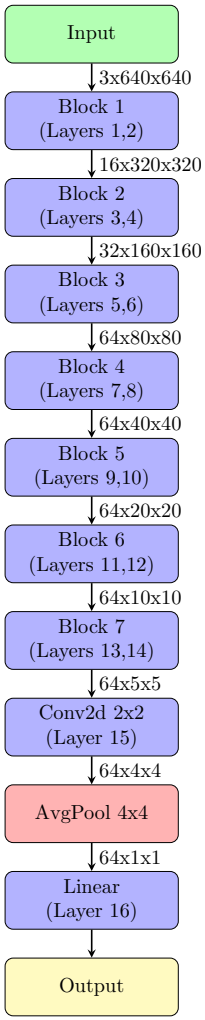
Guidelines

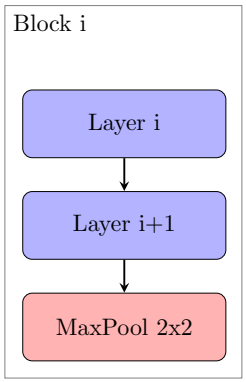
- Use an optimal image resolution
  Tested 128, 256, 384, 512, 640, 724, 768, 892. Optimal: 640
- Use all available information
- Use a fully convolutional neural network
- Use small size convolutions
  Feature extraction 3x3 and 2x2 in classification layer
- Adapt convolution sizes and number of layers to get a RF as close as possible to the image size
- Use ReLU as activation function
- Use batch normalization in every layer
- Use QWK as loss function
- Use a linear classifier
- Use an efficient number of features
  Tested: 32 to 512. Optimal: 64

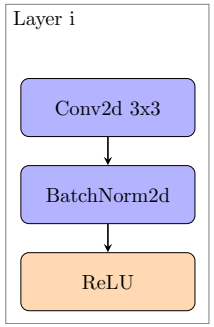
Classification - Enhanced Models

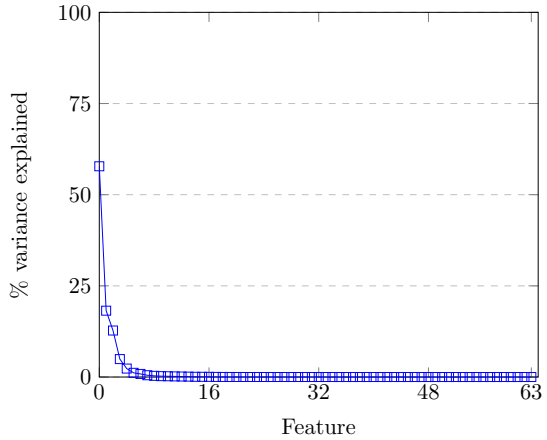
Design

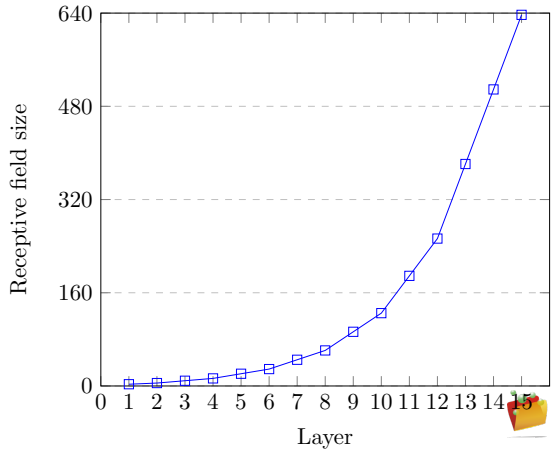
Prediction model of 391,325 trainable parameters located in blue blocks.











Classification - Enhanced Models

Results: EyePACS Dataset over a test set of 10,000 images of 5,000 different patients

Inter-rater agreement 5 classes:

- $QWK = 0.801$ with information of one eye
- $QWK = 0.844$ with information of both eyes (same feature extractor, retrained last linear layer)

Group detection of the most severe cases of DR (classes 2, 3 and 4):

- Sensitivity=0.906 (95% CI: 0.893 to 0.919)
- Specificity=0.847 (95% CI: 0.840-0.855)
- Accuracy=0.857
- $F_1 = 0.710$
- MCC=0.648

Classification - Enhanced Models

Results: Messidor-2 Dataset over a test set of 1,748 images

Inter-rater agreement 4 classes:

- $QWK = 0.830$ with information of one eye

Group detection of the most severe cases of DR (classes 2, 3):

- Sensitivity=0.908 (95% CI: 0.883-0.933)
- Specificity=0.911 (95% CI: 0.890-0.933)
- Accuracy=0.910
- $F_1 = 0.896$
- MCC=0.817

Classification - Enhanced Models

Results: Classification Benchmarks over Messidor-2 Dataset

Reference	Parameters	Depth	Sensitivity	Specificity
(Gulshan et al., 2016)	23,851,784	159	96.1 %	93.9 %
Our work	391,325	17	91.1 %	90.8 %

Table: Prediction performance & model complexity comparison of our proposal vs the state-of-the-art model (Messidor-2 data set)

- Our model differentiate between the five disease classes, detecting also the milder cases (of medical interest for early detection).
- (Gulshan et al., 2016) is a binary classifier specialized in detection of the most severe cases of the disease

Outline

Classification Model Stability

- Deep learning models have a huge parameter set (millions of parameters), ie. are difficult to analyze
- Model validity is tested against a test set ideally coming from the same data distribution, having a statistical information about its general behavior
- For a better understanding of model capabilities and limitations a study of variation of outputs vs variations in inputs is proposed.
- Proposed variables of study: rotation, hue, saturation and lightness

Classification Model Stability

Variables of study: rotation, hue, saturation and lightness

- Rotation
- Hue referring to the attribute of a visual sensation according to which an area appears to be similar to one of the perceived colors: red, yellow, green, and blue, or to a combination of two of them
- Lightness representing the brightness relative to the brightness of a similarly illuminated white
- Saturation showing the colorfulness of a stimulus relative to its own brightness

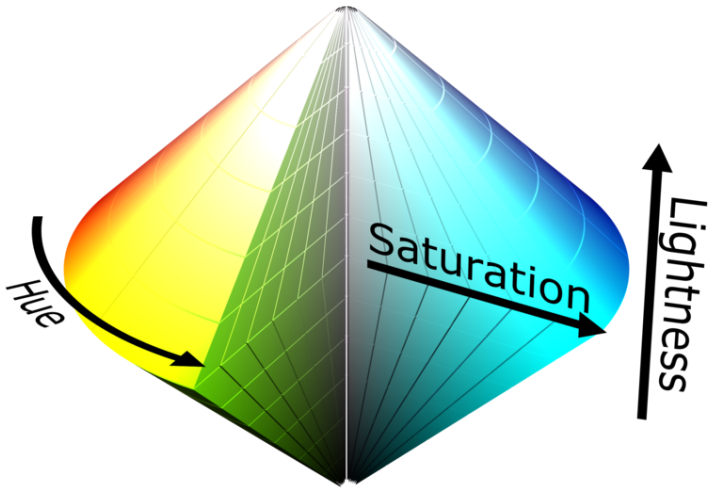
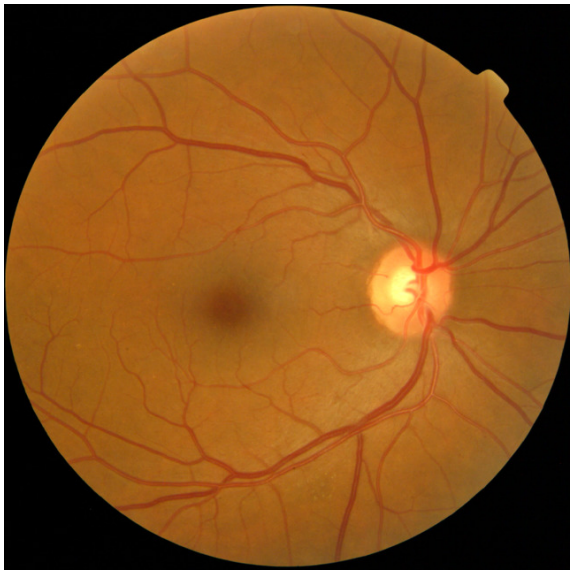
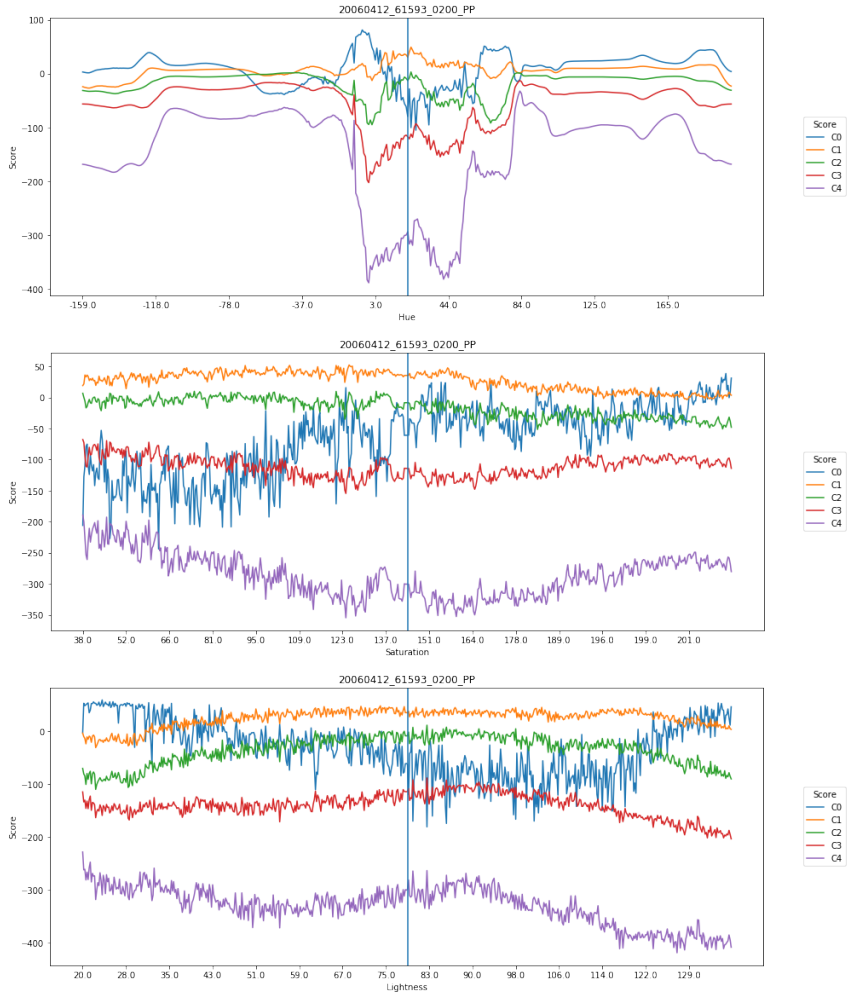


Figure: HSL color space

Stability analysis sample

20060412 61593 0200 PP Target: 1 HSL

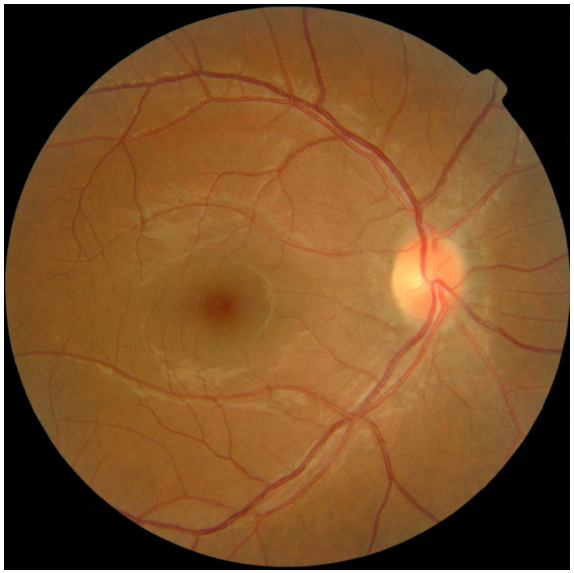


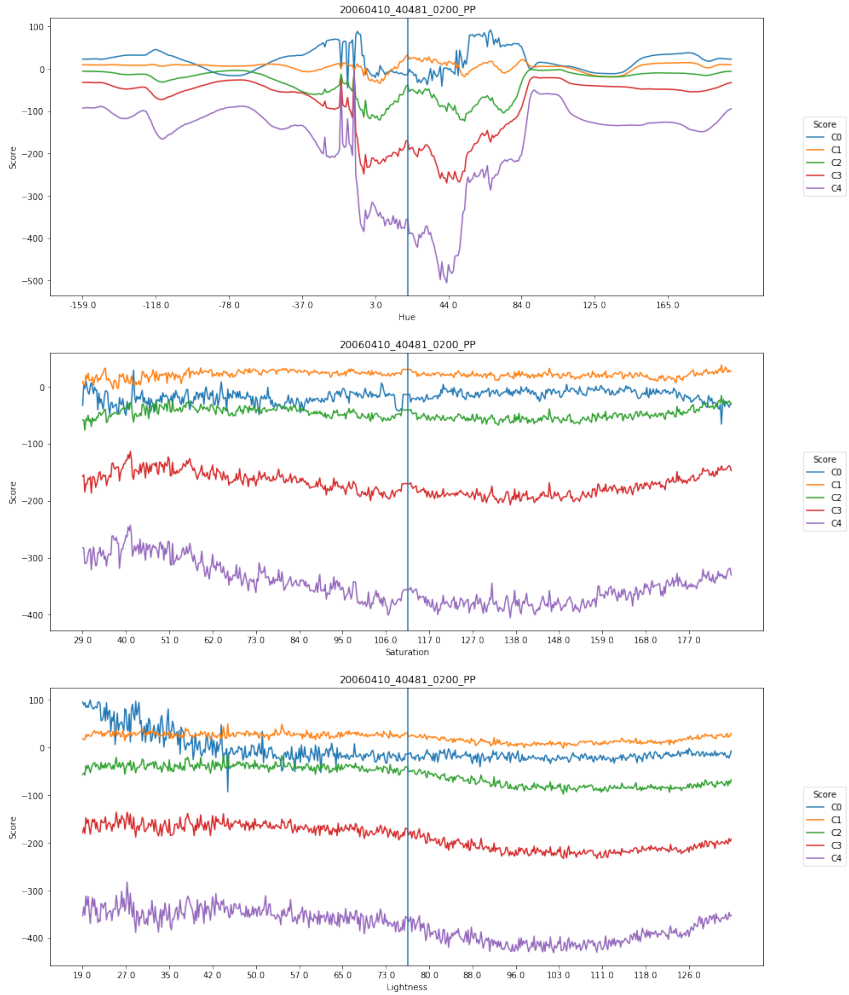


Hue — Saturation — Lightness

Stability analysis sample

20060410 40481 0200 PP Target: 2 HSL

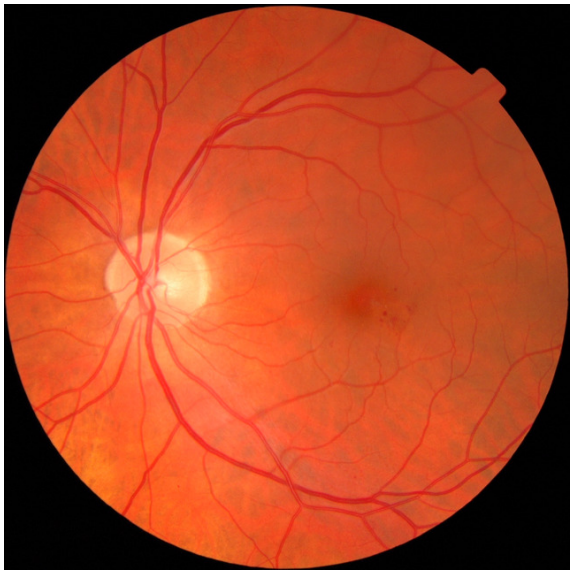


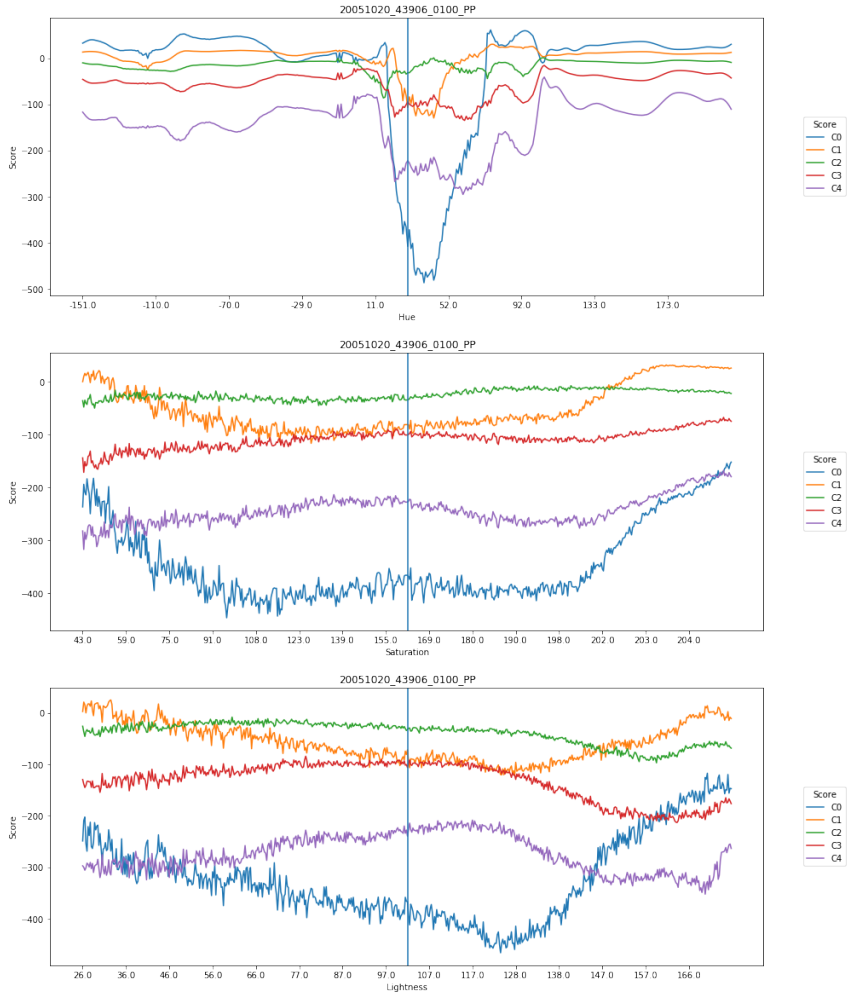


Hue — Saturation — Lightness

Stability analysis sample

20051020 43906 0100 PP Target: 3 HSL

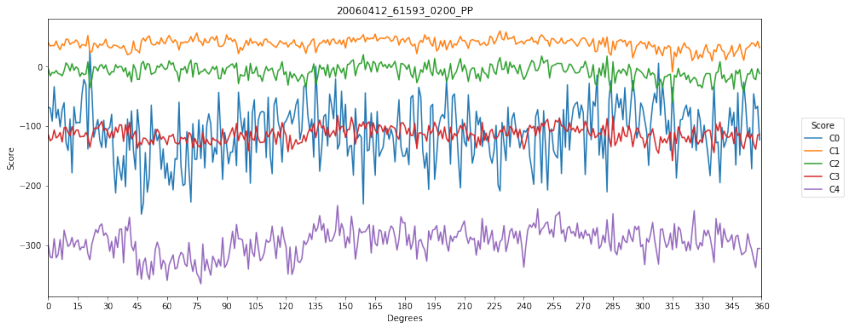




Hue — Saturation — Lightness

Stability analysis sample

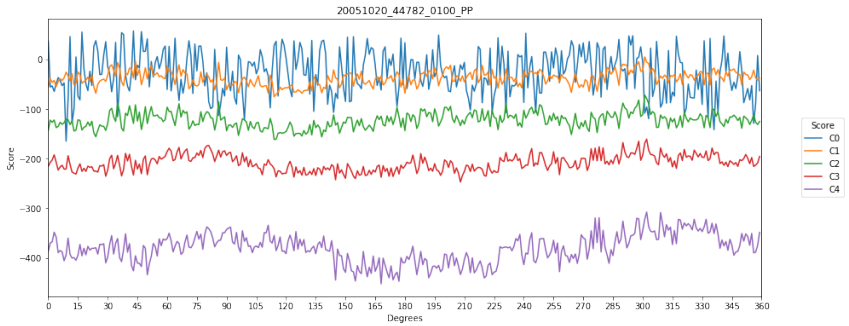
20060412 61593 0200 PP Target: 1 Rotation



Rotation Visualization

Stability analysis sample

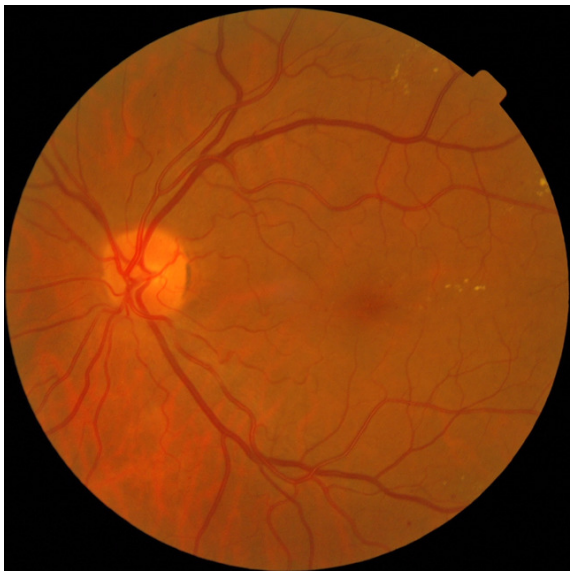
20051020 44782 0100 PP Target: 1 Rotation

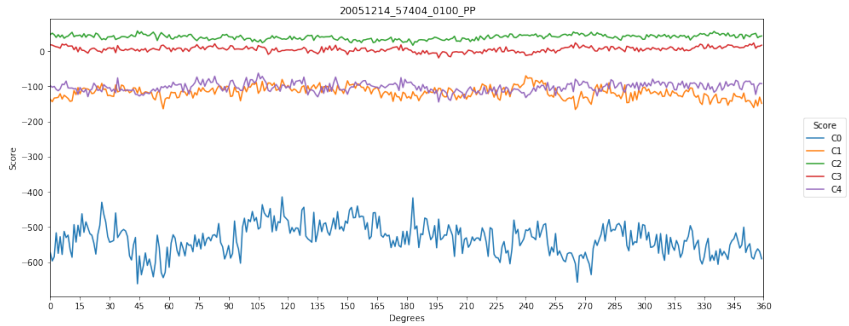


Rotation Visualization

Stability analysis sample

20051214 57404 0100 PP Target: 2 Rotation

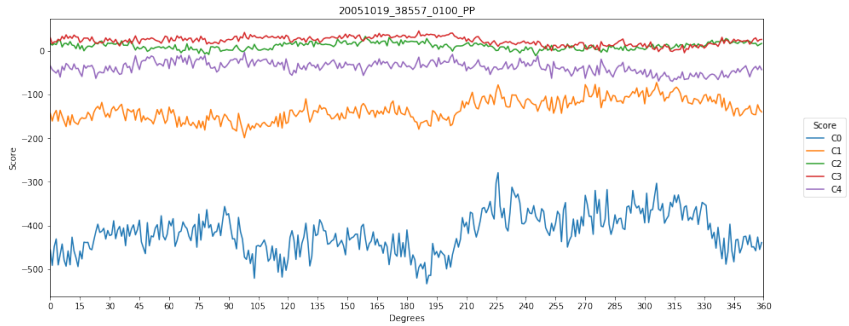




Rotation Visualization

Stability analysis sample

20051019 38557 0100 PP Target: 3 Rotation



Rotation Visualization

Stability Conclusions

- The model is stable under changes in image saturation, lightness and rotation in most of the cases.
- Hue camera calibration is critical for obtaining correct results.
- A full stability study can help improving diagnostic but requires more time.

Part III

Interpretation

Outline

Interpretation

Objective

Interpretation of the outputs reported by the model.

1. Treat model classification outputs as scores
2. Backpropagate them layer by layer through activated nodes until reaching input space
3. In every layer a score map distribution of the correspondent receptive field is obtained
4. Doing such backpropagation until reaching input space a input-space score map distribution is obtained

Interpretation

Model

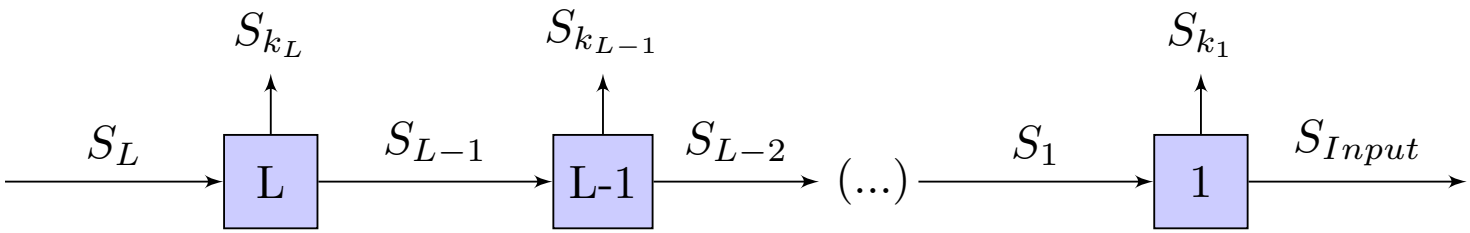
Proposition 1:

$$S_l = \lambda_l a_l$$

Proposition 2:

$$S_{l+1} = S_l + S_{k_l}$$

Score propagation scheme:



Score conservation equation:

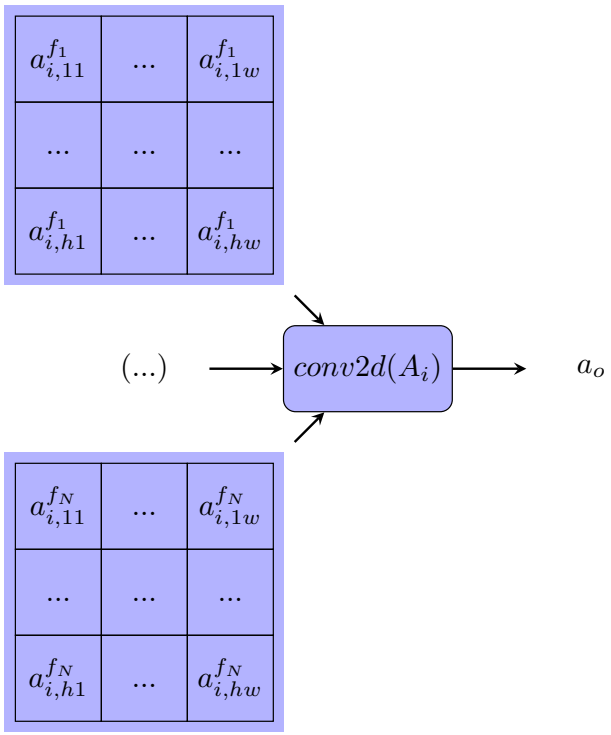
$$S_L = \sum_{l=1}^{L} \left(\sum S_{k_l} \right) + \left(\sum S_{Input} \right)$$

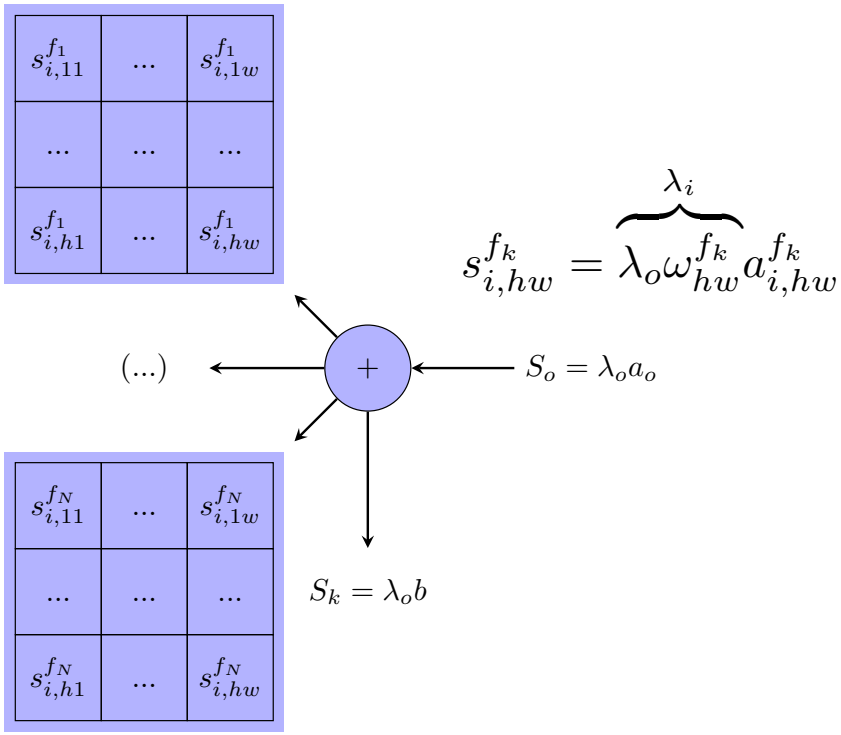
- Such two propositions are enough for the derivation of a general method of score propagation.
- For each one of the typical deep learning blocks a score propagation model is derived.

Interpretation

Derivation of score propagation model

Score propagation through a convolutional layer:

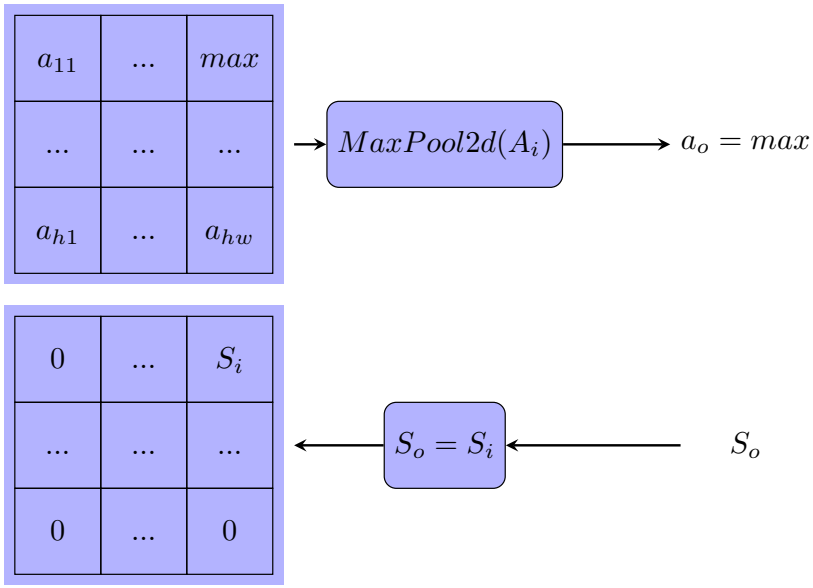




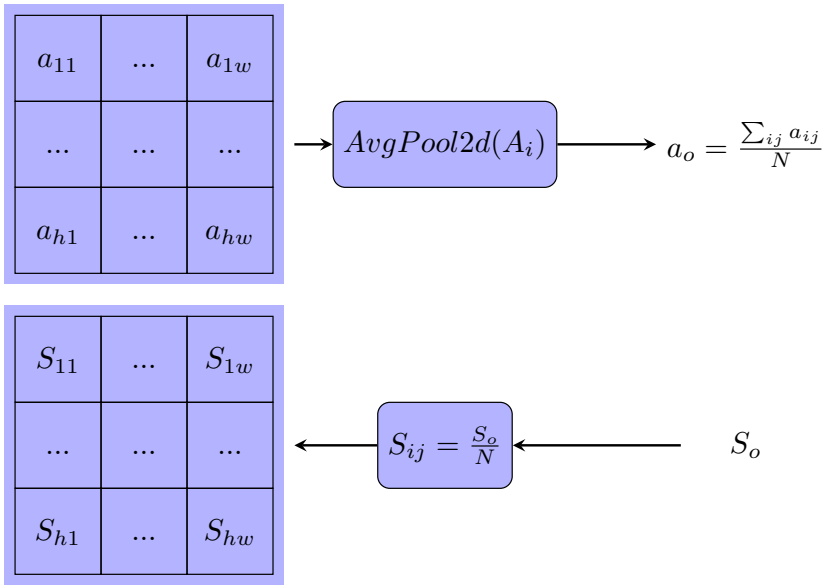
Interpretation

Derivation of score propagation model

Max pooling:



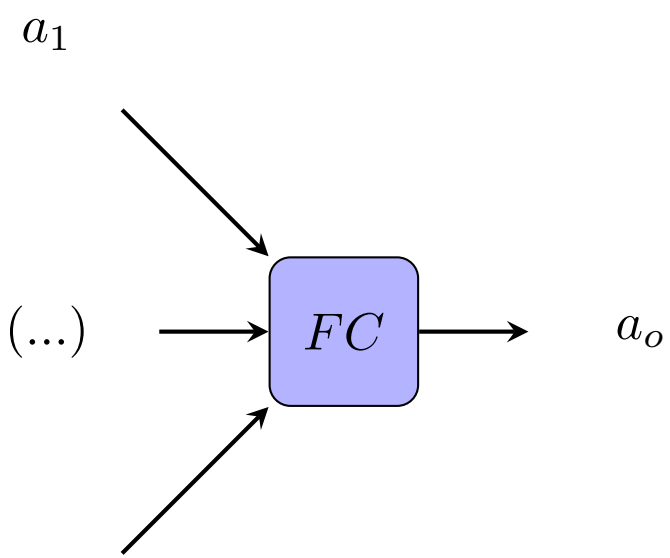
Average pooling:

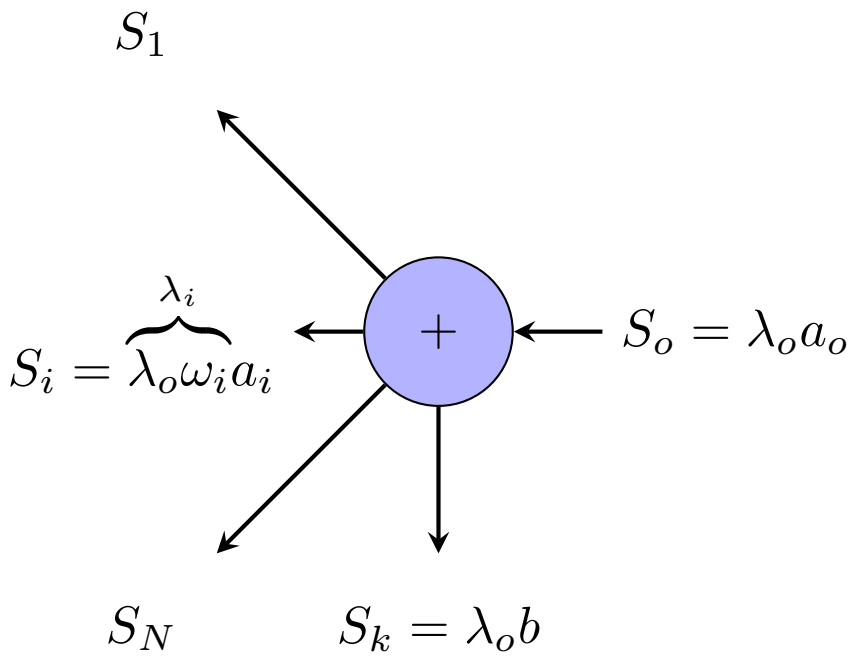


Interpretation

Derivation of score propagation model

Fully connected layer:

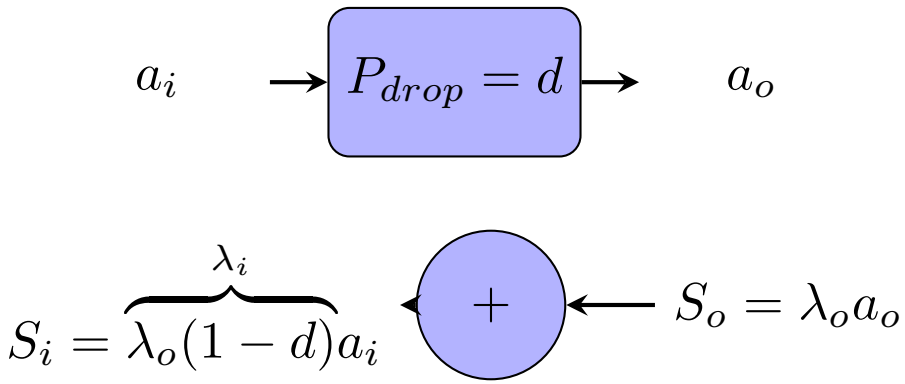




Interpretation

Derivation of score propagation model

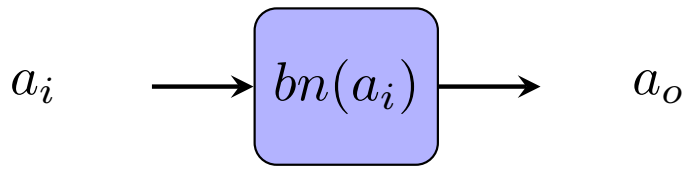
Dropout layer:

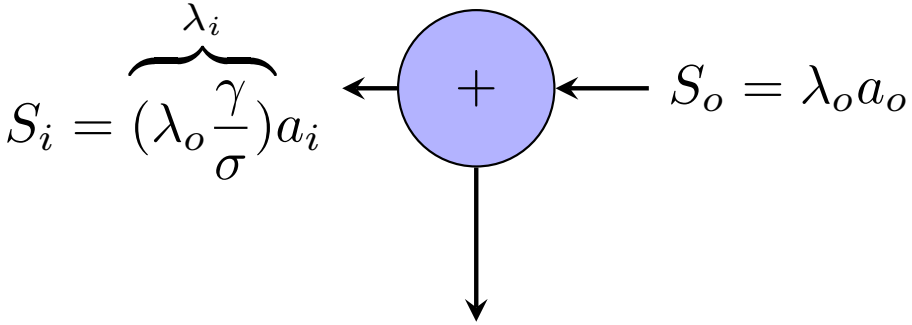


Interpretation

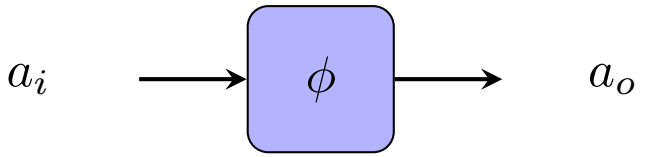
Derivation of score propagation model

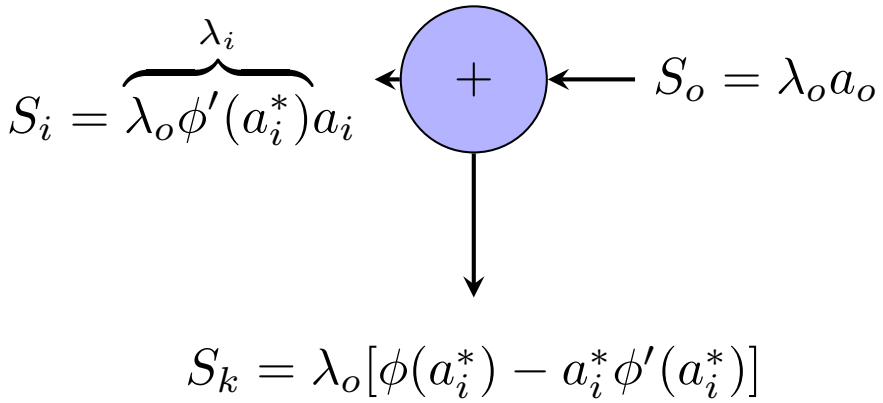
Score propagation through a batch normalization node:





Score propagation through an activation function node:





Interpretation

Mapping the score of hidden layers and S_k to input-space

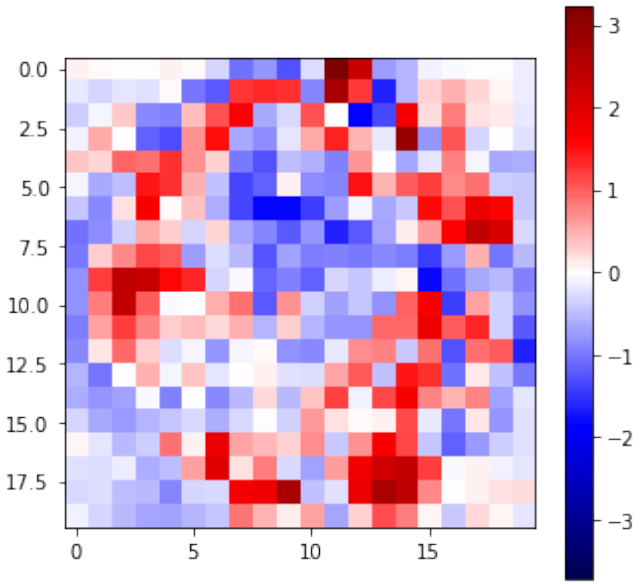
- Every node has two score constituents: one input-dependent, that can be easily forwarded, and another one RF-dependent, i.e layer-dependent.
- Effective RF is not equal to the theoretical RF (Luo et al., 2016). The effective one acts more like a 2D-gaussian function, where the points located in the borders contribute less than the center ones.
- Using such prior information, it is possible to make an approximate conversion of the full and constant scores in the hidden-space to the input-space using a 2D-gaussian prior.

Interpretation

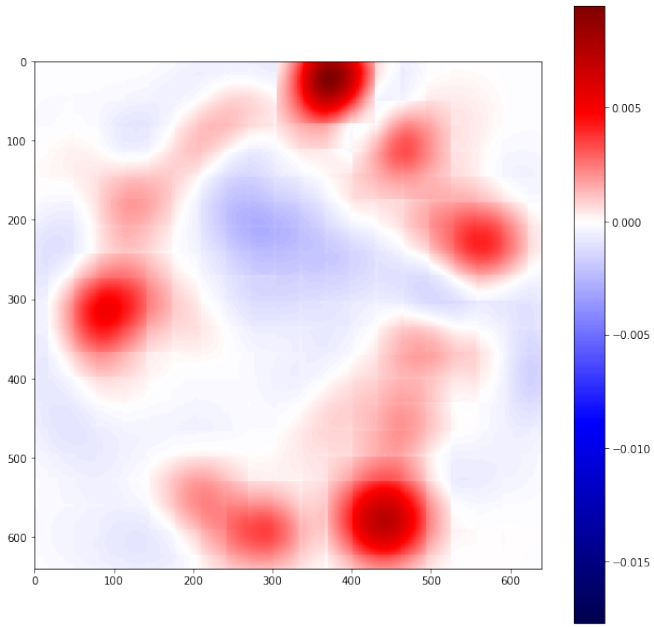
Mapping the score of hidden layers and S_k to input-space (example)

For example, for a 20x20 hidden layer with a RF of 125x125 pixels, each point represents the cummulative value of a gaussian distribution. Summing up each gaussian distribution we obtain the map in input space.

Hidden layer activations, RF=125^2:



Mapped to input-space:



Interpretation

Samples

- Class 4 score feature-wise visualization propagation
- Class 4 score layer-wise visualization propagation

- Class 1: 11736 left
- Class 4: 11854 left
- Class 3: 1561 left
- Class 0: 162 right
- Class 2: 20051019 38557
- Class 2: 20051020 43906
- Class 2: 20051021 52127
- Class 2: 20051214 57404
- Class 2: 20051216 44939
- Class 1: 20060410 40481
- Class 1: 20060411 62228
- Class 1: 20060412 59658
- Class 2: 20060412 59717
- Class 1: 20060412 61593
- Class 0: 20060523 45524
- Class 2: 20060523 50392

Interpretation

Conclusions

- We designed a method for deep learning models result interpretation based on input-space score map generation
- The model is designed for general applicability in different domains and for different networks
- We show an application based on interpretation of the results reported in our diabetic retinopathy disease grading model.
- The interpretation model is able to identify correctly the lesions present in images.
- Such identification is inferred only from the information coming from the disease grading labels of the training set
- This results prove, not only that the interpretation model is successful in determining the causes under a particular classification but also that the original model is able to identify the important features of the image, ie. the correct statistical regularities.

Outline

Feature Space Compression

- Neural networks feature spaces are frequently high-dimensional with high correlation values between dimensions, ie. are low dimensional manifolds embedded in high dimensional spaces.
- We present a methodology for linear compression of the feature space allowing the feature space compression from the original 64 dimensions to only 3 with a reduction of performance lower than 2.5%.
- This feature-space compression aims to facilitate the interpretation.

Feature Space Compression

Methodology

Initial Classifier

Modified Classifier

Feature Space Compression

Mathematical Formalization

$$F_{train} = \{\boldsymbol{f^{(i)}} : i = 1..T\}, \quad \boldsymbol{f^{(i)}} = (f_1^{(i)}, f_2^{(i)}, ...f_m^{(i)}) \tag{1}$$

$$S_{train} = \{\boldsymbol{s^{(i)}} : i = 1..T\}, \quad \boldsymbol{s^{(i)}} = (s_1^{(i)}, s_2^{(i)}, ...s_n^{(i)}) \tag{2}$$

$$\boldsymbol{s^{(i)}} = \boldsymbol{W}\boldsymbol{f^{(i)}} \tag{3}$$

$$\max_{\boldsymbol{A}} \left[\kappa_{val}(C_{train})\right] \tag{4}$$

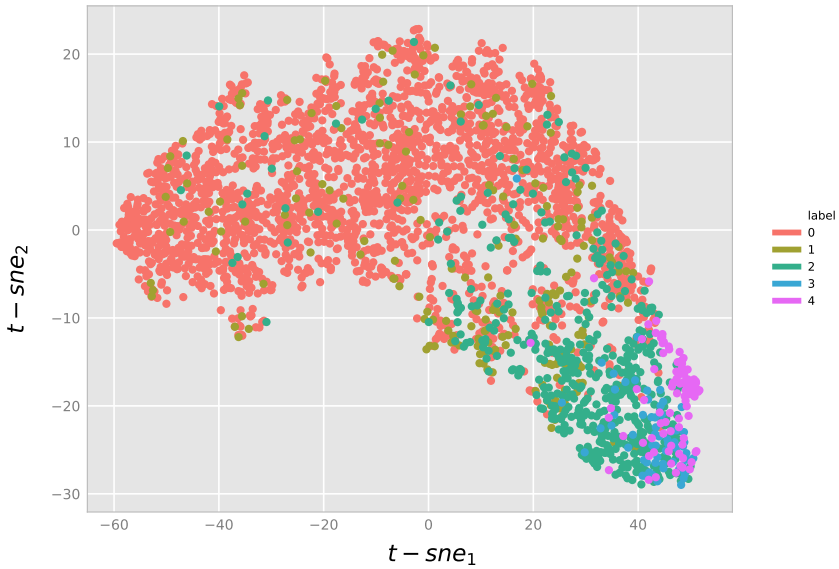
$$C_{train} = \{\boldsymbol{A}\boldsymbol{f^{(i)}}, \forall \boldsymbol{f^{(i)}} \in F_{train}\} \tag{5}$$

$$C'_{train} = \{\boldsymbol{B}\boldsymbol{s^{(i)}}, \forall \boldsymbol{s^{(i)}} \in S_{train}\} \tag{6}$$

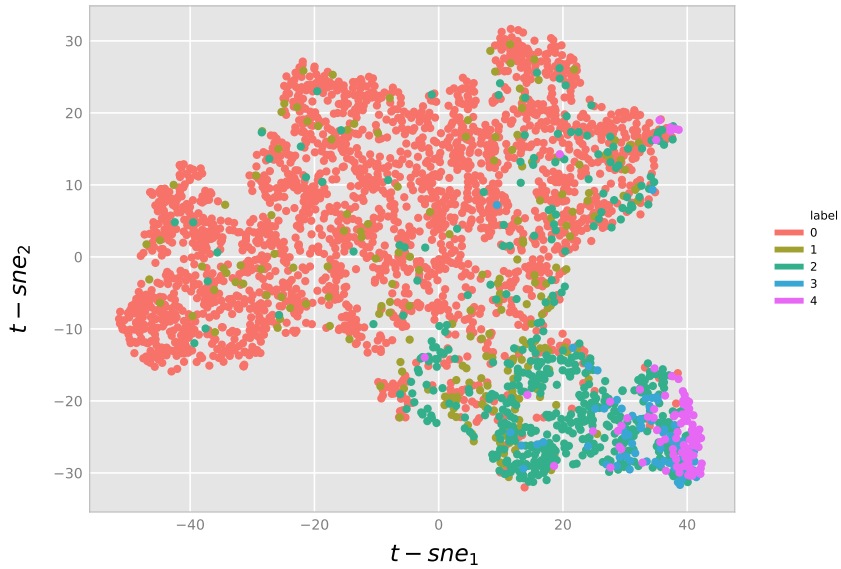
$$\min_{n} \left[\kappa_{val}(C'_{train}) - \kappa_{val}(C_{train})\right] \tag{7}$$

Feature Space Compression

Comparison between the 2D t-SNE visualization of validation set using the original feature space and the final 3-dimensional ICA space:



(a) Original feature space (64 comp.)



(b) ICA space (3 comp.)

Feature space compression

Conclusions

- We designed a method for the internal compression of the feature space model representation
- The method is of general applicability to other networks and applications
- In the diabetic retinopathy disease grading case the method allows the compression of the original 64 features internal representation into only 3 features with a loss of perfomance lower than 2.5%.
- Reducing the number of features and the correlation between them, facilitates its interpretation by human experts.

Part IV

Application and Conclusions

Outline

Experimental Application

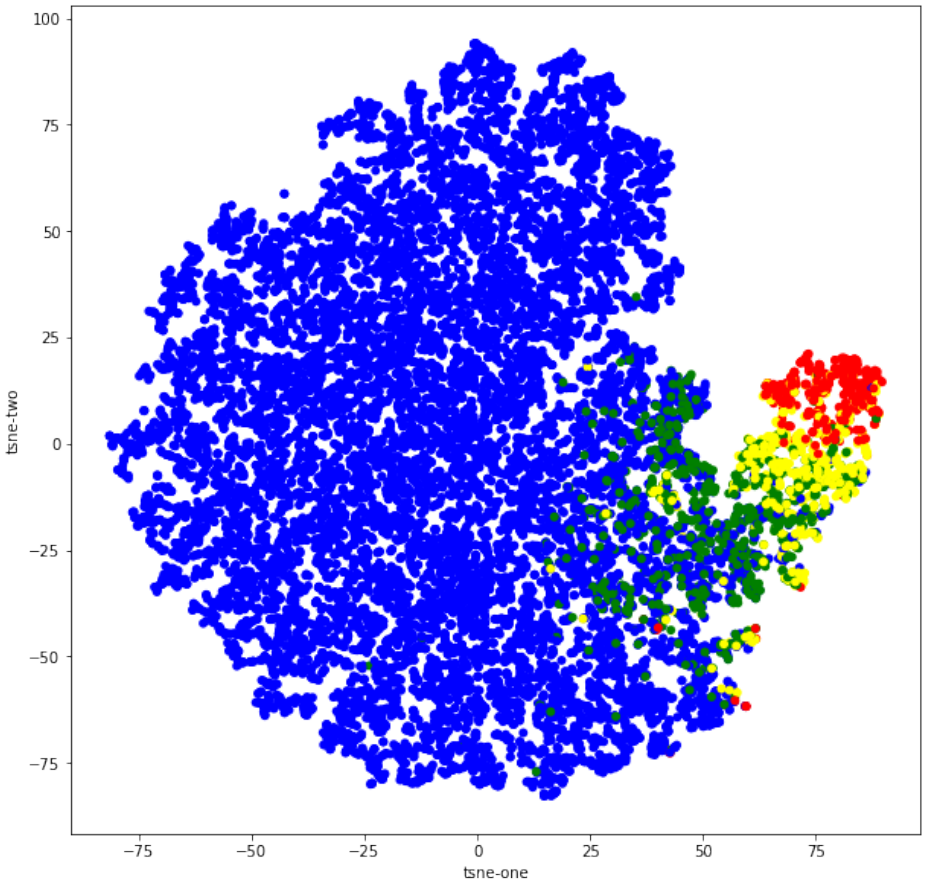
Inference using HUSJR data

- The objective of this study is determine the applicability of the designed model for the prediction in a real case of the Hospital Universitari Sant Joan de Reus.
- For this purpose a database of $19,230$ tagged retinographies is used.
- Some discrepancies are detected in the definition of the classes used in HUSJR and the ones of the trained model.
- After an initial evaluation with original model a retraining is applied for making compatible the class definitions.
- Messidor-2 class definition is defined as the new standard.
- Last linear classification layer is retrained.

Experimental Application

Inference using HUSJR data

t-SNE visualization of the feature space representation of samples of HUSJR dataset. Blue (class 0), green (class 1), yellow (class 2) and red (class 3)



Experimental Application

Inference using HUSJR data

HUSJR confusion matrix using EyePACS trained original model. $QWK = 0.791$

	Pred 0	Pred 1	Pred 2	Pred 3	Pred 4
True 0	15,112	2,024	159	6	12
True 1	11	547	326	14	3
True 2	0	4	439	133	7
True 3	0	1	20	358	54

HUSJR confusion matrix with EyePACS trained original model plus a linear classifier retrained using Messidor-2 Dataset. $QWK = 0.823$

	Pred 0	Pred 1	Pred 2	Pred 3
True 0	15,277	1,944	83	9
True 1	5	595	284	17
True 2	0	2	456	125
True 3	0	1	6	426

Experimental Application

Inference using HUSJR data

The indexes obtained for classification of the most severe cases of the disease (considering positive class = 2,3 and negative class = 0,1) are:

- Sensitivity = 0.997
- Specificity = 0.978
- Positive predictive value (PPV) = 0.720
- Negative predictive value (NPV) = 0.9998
- Accuracy (ACC) = 0.979
- F_1 Score = 0.836

Experimental Application

Conclusions of inference using HUSJR data

- We studied the applicability of our best model for the prediction of diabetic retinopathy, trained using the EyePACS dataset, for the prediction of diabetic retinopathy in HUSJR
- A feature space visualization has been done in order to check its capacity for separating between classes.
- A first evaluation of the model predictability was done, obtaining good results, but with small loss in performance. This loss was probably produced to the slight differences in class definitions.
- After a class standardization, the linear classifier of the model was retrained using Messidor-2 Dataset, using as a feature extractor the original model.
- Performance of the new model was tested again, reaching values of inter-rater agreement similar to the obtained by expert ophthalmologists.
- Medical team is extremely satisfied with obtained results.

Outline

Summary of contributions I

The main contributions of this thesis are:

1. Design of automatic classifiers based on deep neural networks able to reach ophthalmologist performance level.
   Jordi de la Torre, Aïda Valls, and Domenec Puig (2016). "Diabetic Retinopathy Detection Through Image Analysis Using Deep Convolutional Neural Networks". In: *Artificial Intelligence Research and Development - Proceedings of the 19th International Conference of the Catalan Association for Artificial Intelligence, Barcelona, Catalonia, Spain, October 19-21, 2016*. Ed. by Àngela Nebot, Xavier Binefa, and Ramon López de Mántaras. Vol. 288. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 58–63. ISBN: 978-1-61499-695-8

Summary of contributions II

The main contributions of this thesis are:

2. Study of the usage of Quadratic Weighted Kappa index as a Deep Learning Loss Function for the optimization of ordinal regression problems.
   Jordi de la Torre, Domenec Puig, and Aida Valls (2018). "Weighted kappa loss function for multi-class classification of ordinal data in deep learning". In: *Pattern Recognition Letters* 105, pp. 144–154 Impact Factor: 1.952 (Q2)
3. Study of the feature space manifold stability of the designed diabetic retinopathy classifiers.
4. Design of a generalized model for the interpretation of results reported by deep learning classifiers.
   Jordi de la Torre, Aida Valls, and Domenec Puig (2017). "A Deep Learning Interpretable Classifier for Diabetic Retinopathy Disease Grading". In: *arXiv preprint arXiv:1712.08107* Accepted for publication in *Neurocomputing*. Impact Factor: 3.241 (Q1)

Summary of contributions III

The main contributions of this thesis are:

5. Design of a method for compressing feature space internal representations of deep learning models.
   Jordi de La Torre, Aida Valls, and Domenec Puig (2018).
   "Identification and Visualization of the Underlying Independent Causes of the Diagnostic of Diabetic Retinopathy made by a Deep Learning Classifier". In: *CoRR* abs/1809.08567. arXiv: 1809.08567.
   URL: http://arxiv.org/abs/1809.08567
   Under revision in *Computer Methods and Programs in Biomedicine*.
   Impact Factor: 2.674 (Q2)
6. Application of designed classifiers into a real use case in Hospital de Reus. A software has been implemented for DR classification and lesion identification. Registered in Benelux Office for Intellectual Property. Reference number 109999.

Outline

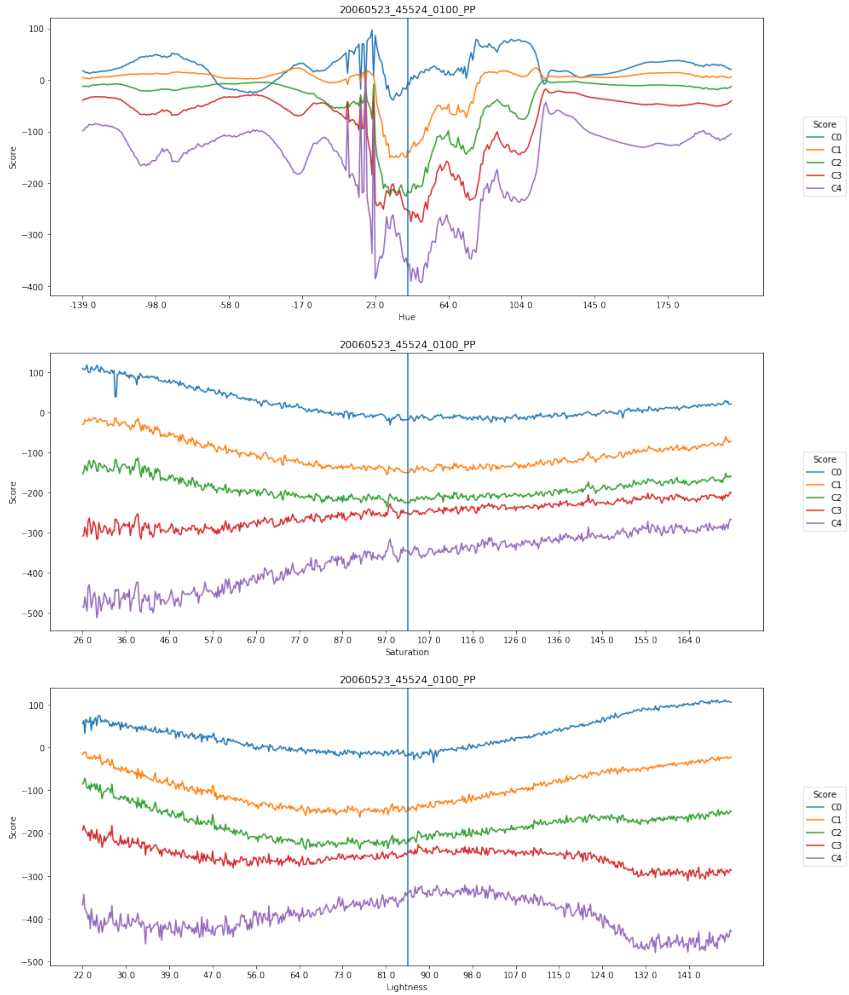
Future possible research lines

- Increase the number of classes to predict
- Transfer learning: Successful networks in challenging tasks like ImageNet, can be good candidates to perform well in specialized medical imaging tasks like ours.
- Unsupervised learning: Generative Adversarial Networks can also be explored for generating new high quality samples from the original dataset
- Reinforcement Learning: Adding to the models the possibility of enhancing its performance, designing online learning methods that allow continuous learning of networks from the corrections done by ophthalmologists on inference time
- Use of Interpretation Model in other applications
- Interpretation model for unsupervised image segmentation

Thank you.

Stability analysis sample

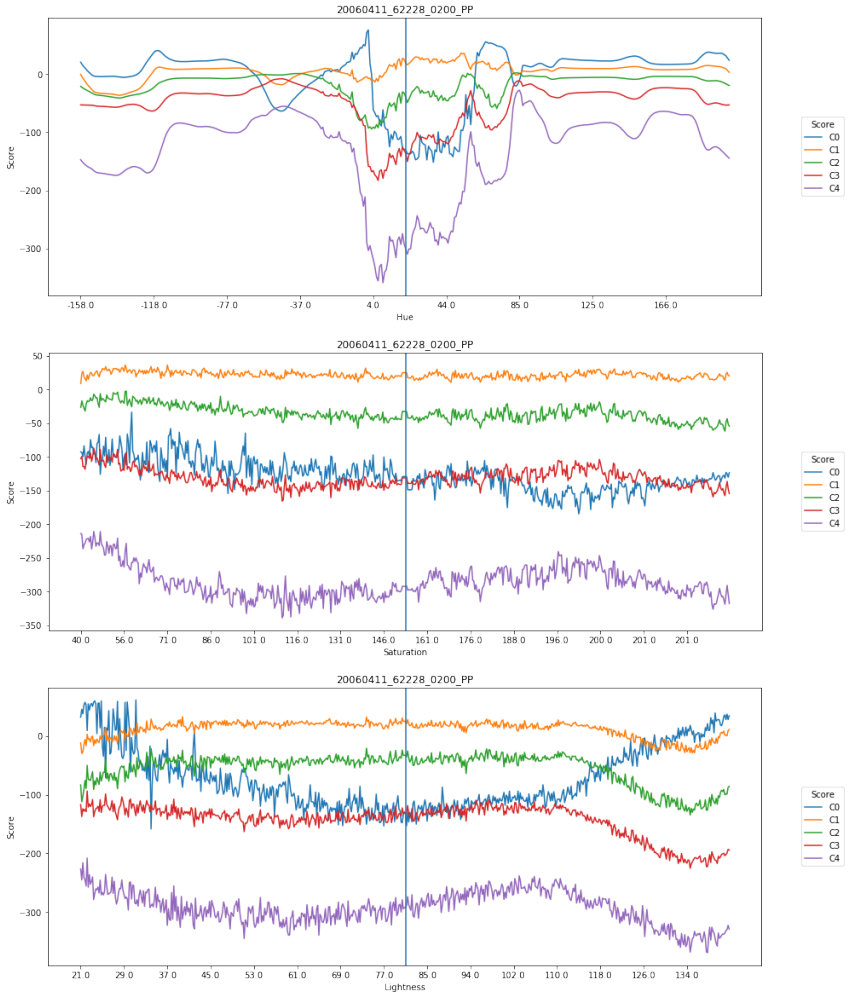
20060523 45524 0100 PP Target: 0 HSL



Hue — Saturation — Lightness

Stability analysis sample

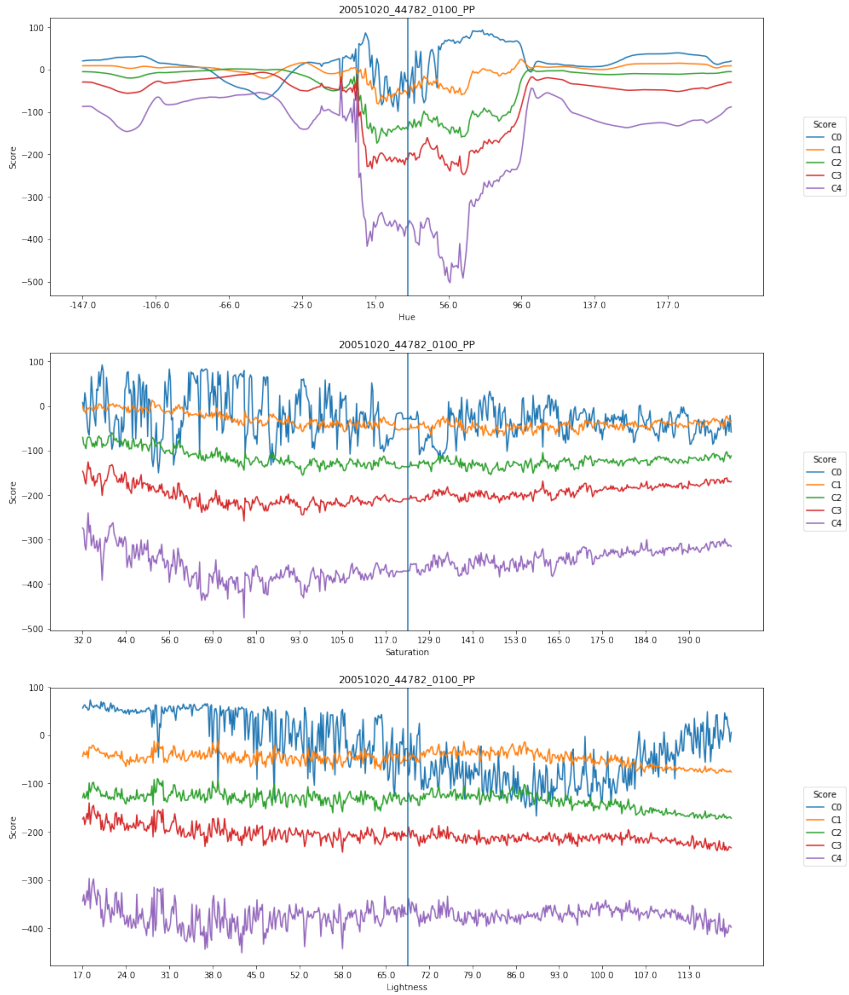
20060411 62228 0200 PP Target: 1 HSL



Hue — Saturation — Lightness

Stability analysis sample

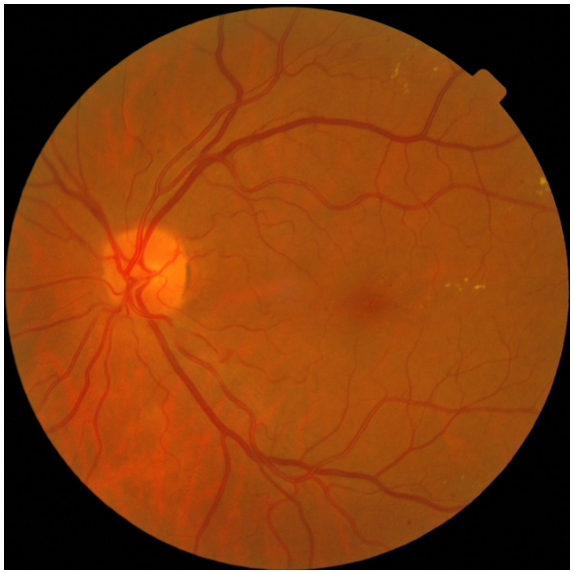
20051020 44782 0100 PP Target: 1 HSL

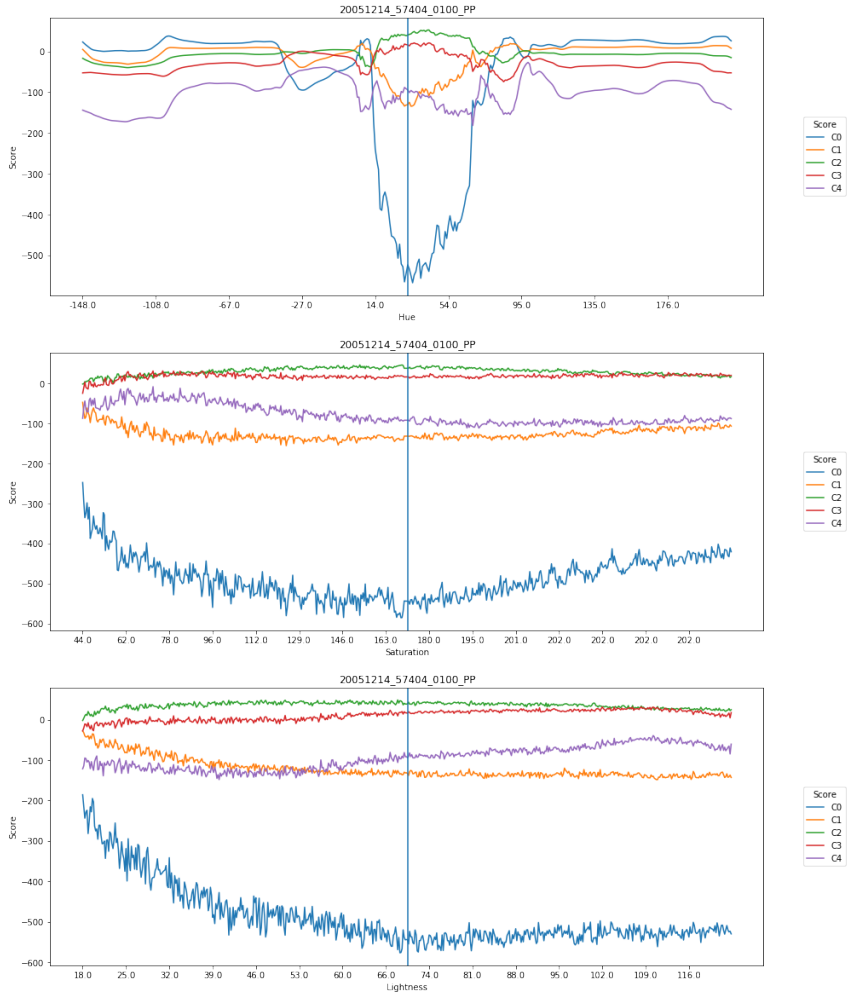


Hue — Saturation — Lightness

Stability analysis sample

20051214 57404 0100 PP Target: 2 HSL

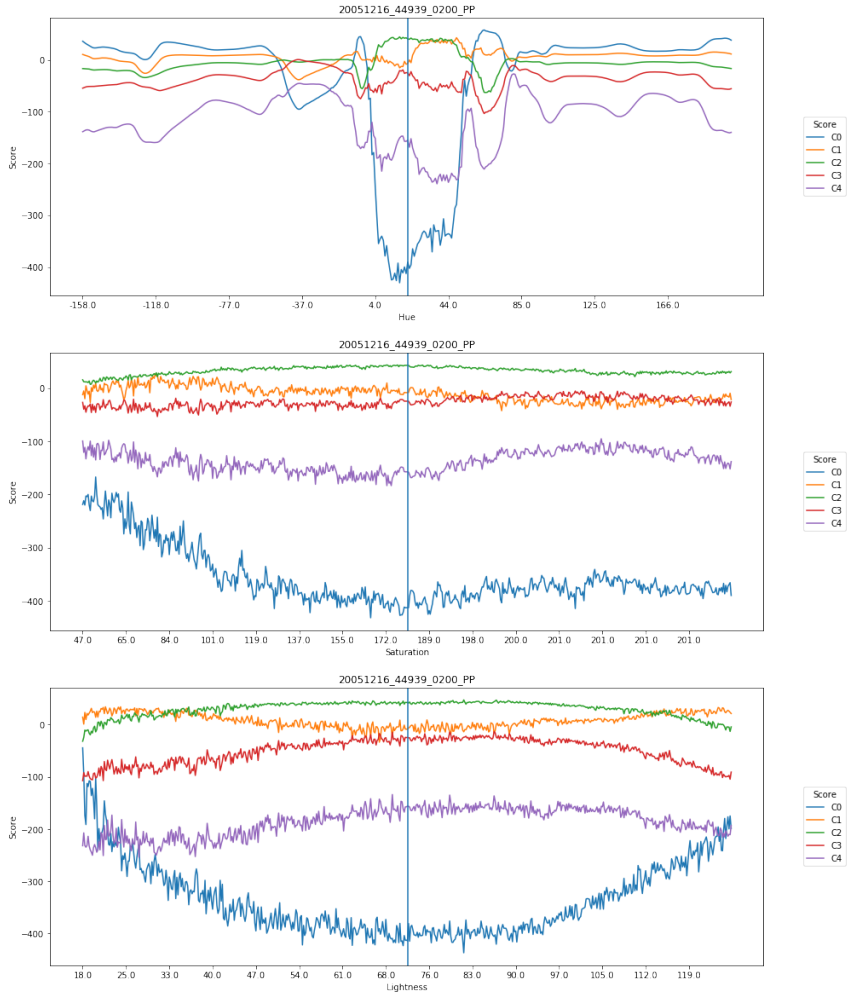




Hue — Saturation — Lightness

Stability analysis sample

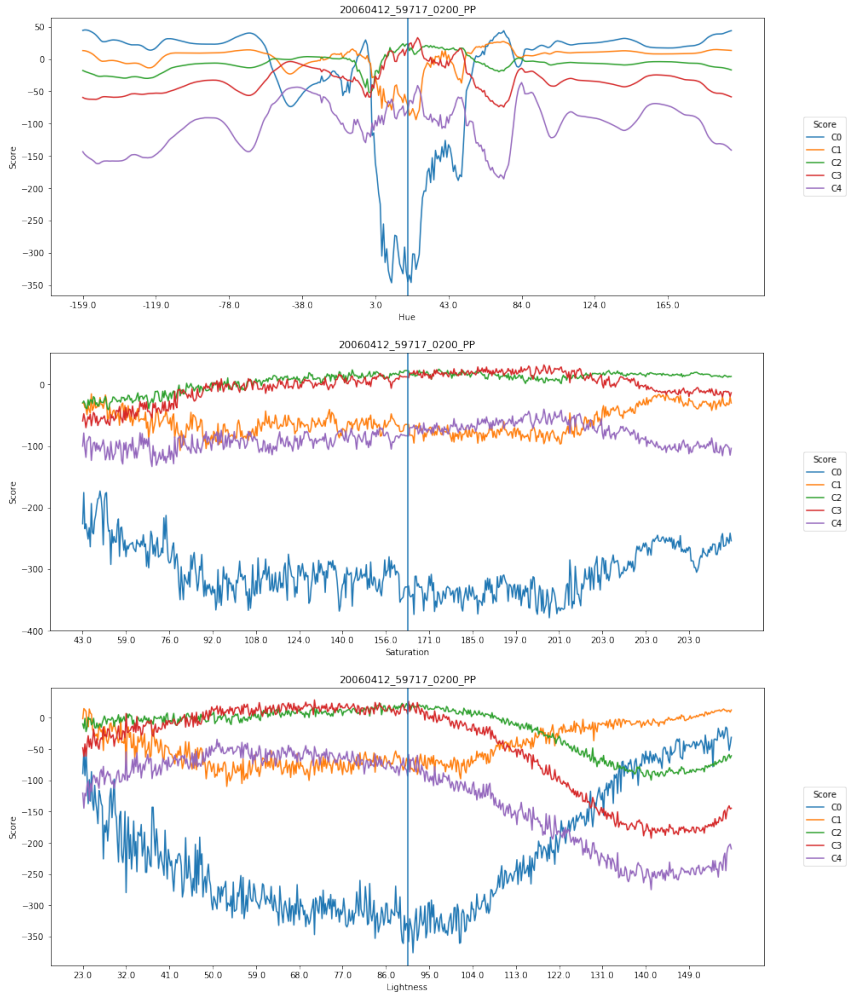
20051216 44939 0200 PP Target: 2 HSL



Hue — Saturation — Lightness

Stability analysis sample

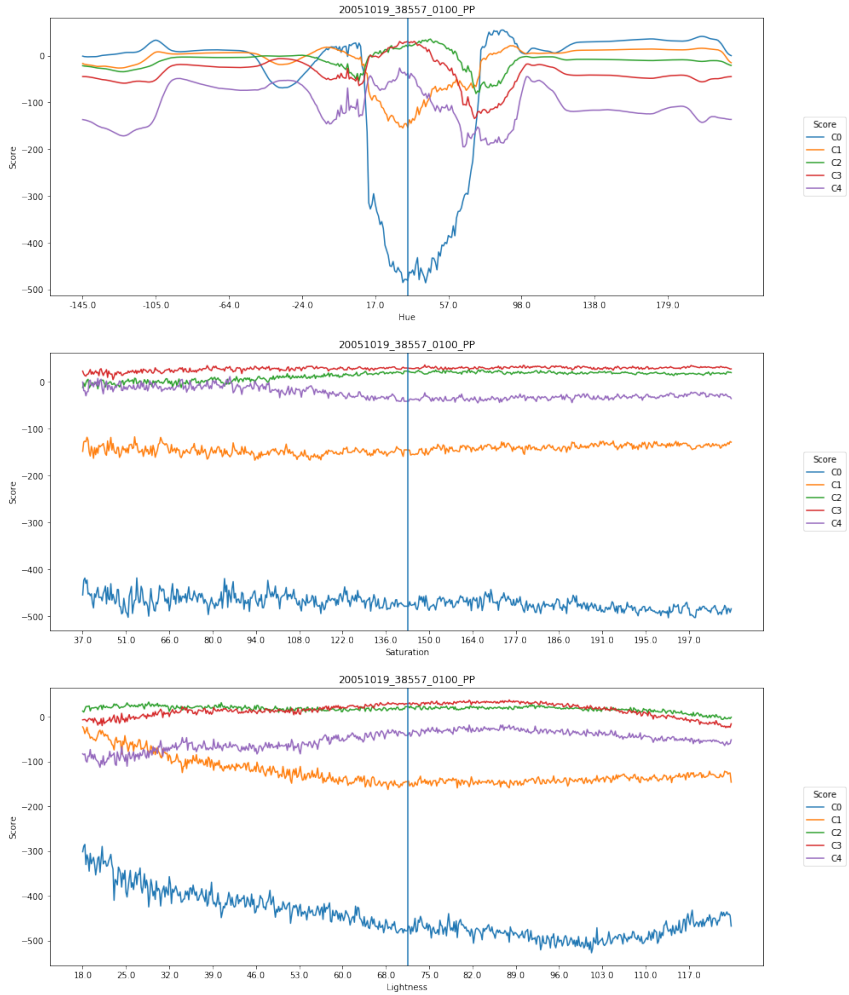
20060412 59717 0200 PP Target: 3 HSL



Hue — Saturation — Lightness

Stability analysis sample

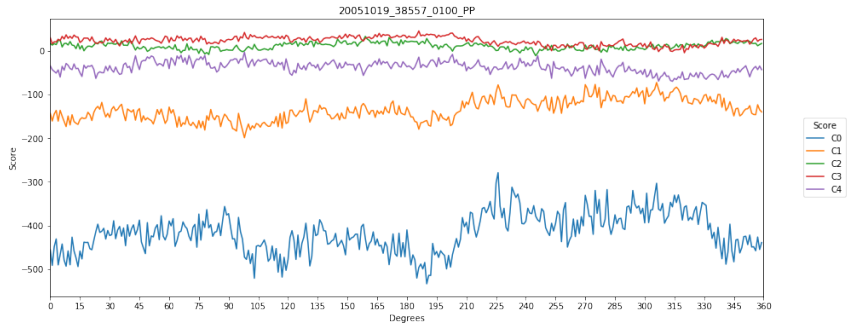
20051019 38557 0100 PP Target: 3 HSL



Hue — Saturation — Lightness

Stability analysis sample

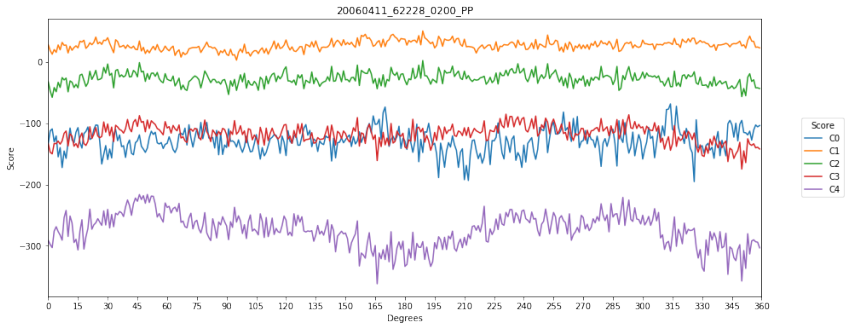
20051019 38557 0100 PP Target: 3 Rotation



Rotation Visualization

Stability analysis sample

20060411 62228 0200 PP Target: 1 Rotation



Rotation Visualization

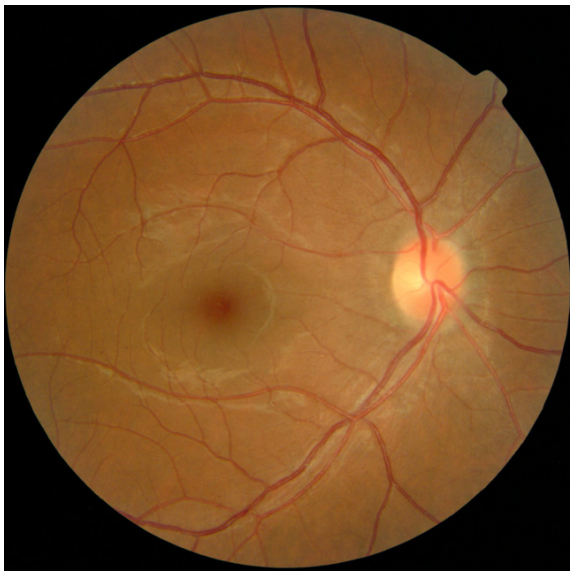
Stability analysis sample

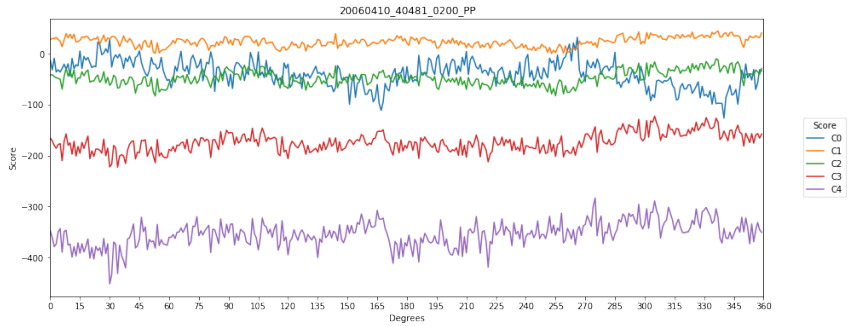
20060412 59658 0200 PP Target: 1 Rotation

Rotation Visualization

Stability analysis sample

20060410 40481 0200 PP Target: 1 Rotation

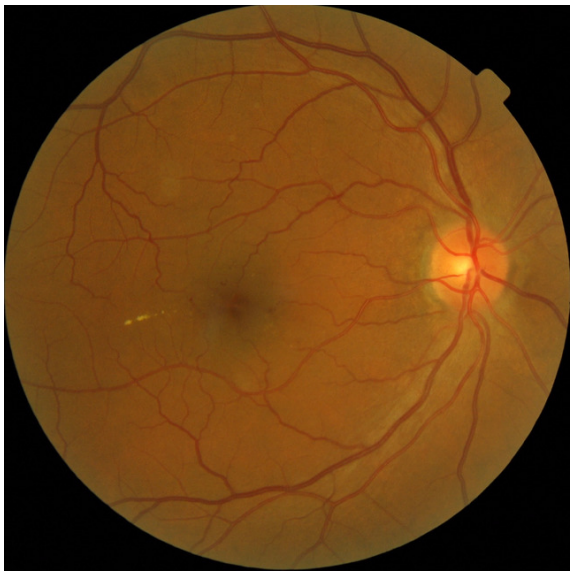


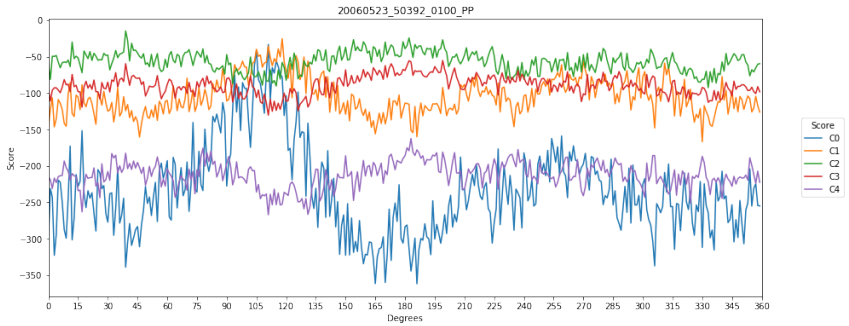


Rotation Visualization

Stability analysis sample

20060523 50392 0100 PP Target: 2 Rotation

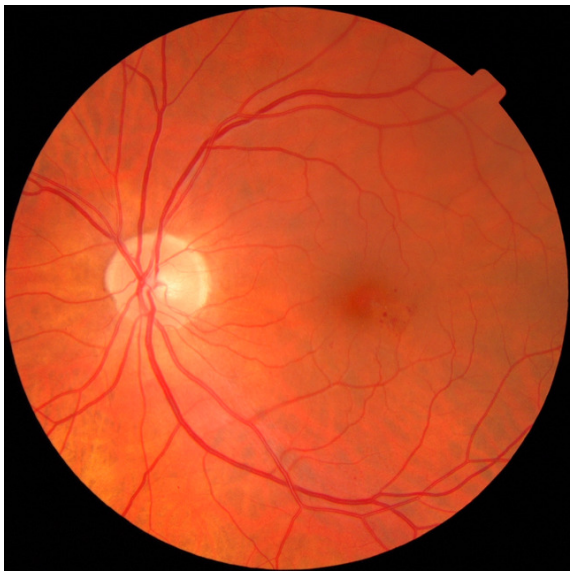


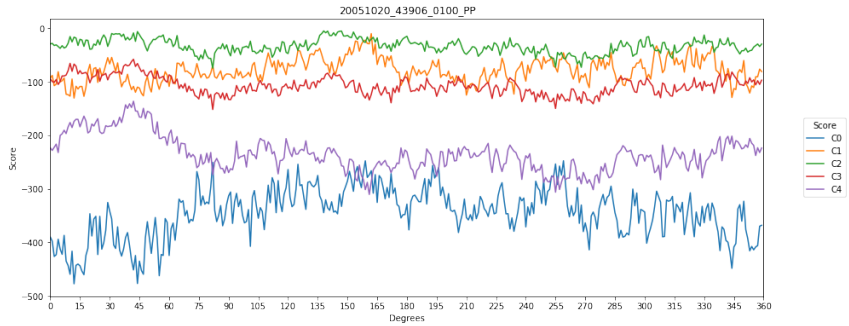


Rotation Visualization

Stability analysis sample

20051020 43906 0100 PP Target: 3 Rotation

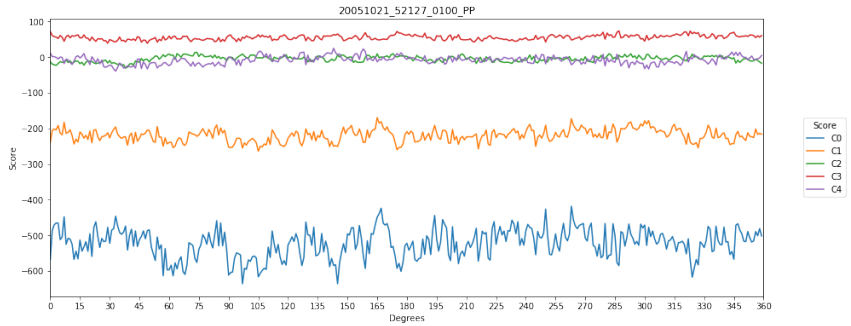




Rotation Visualization

Stability analysis sample

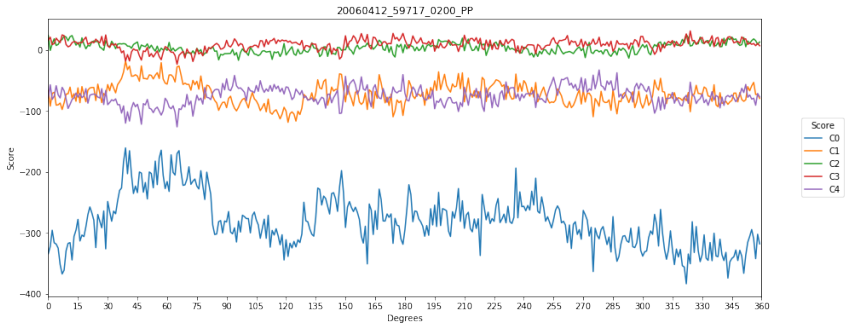
20051021 52127 0100 PP Target: 3 Rotation



Rotation Visualization

Stability analysis sample

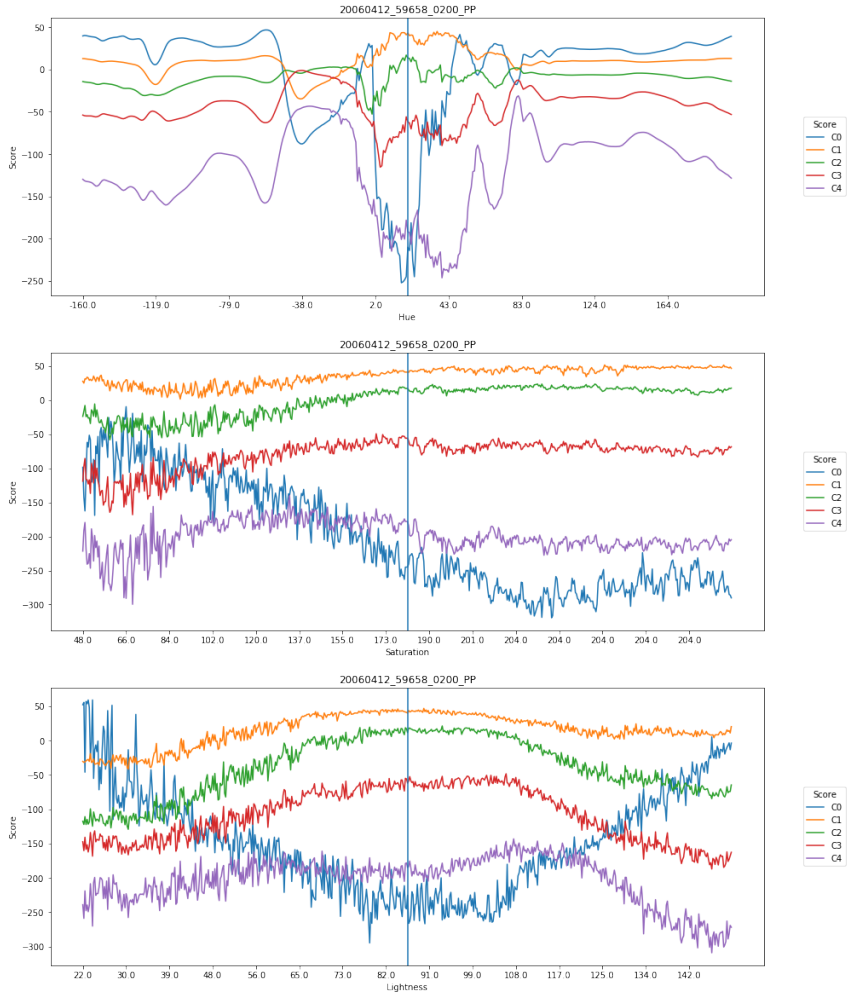
20060412 59717 0200 PP Target: 3 Rotation



Rotation Visualization

Stability analysis sample

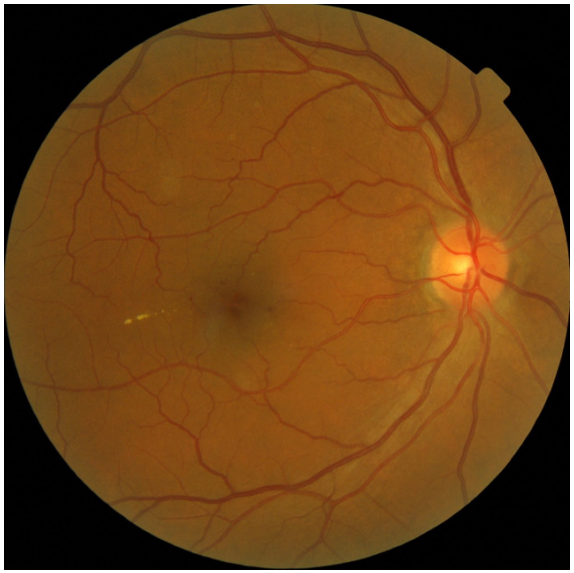
20060412 59658 0200 PP Target: 1 HSL

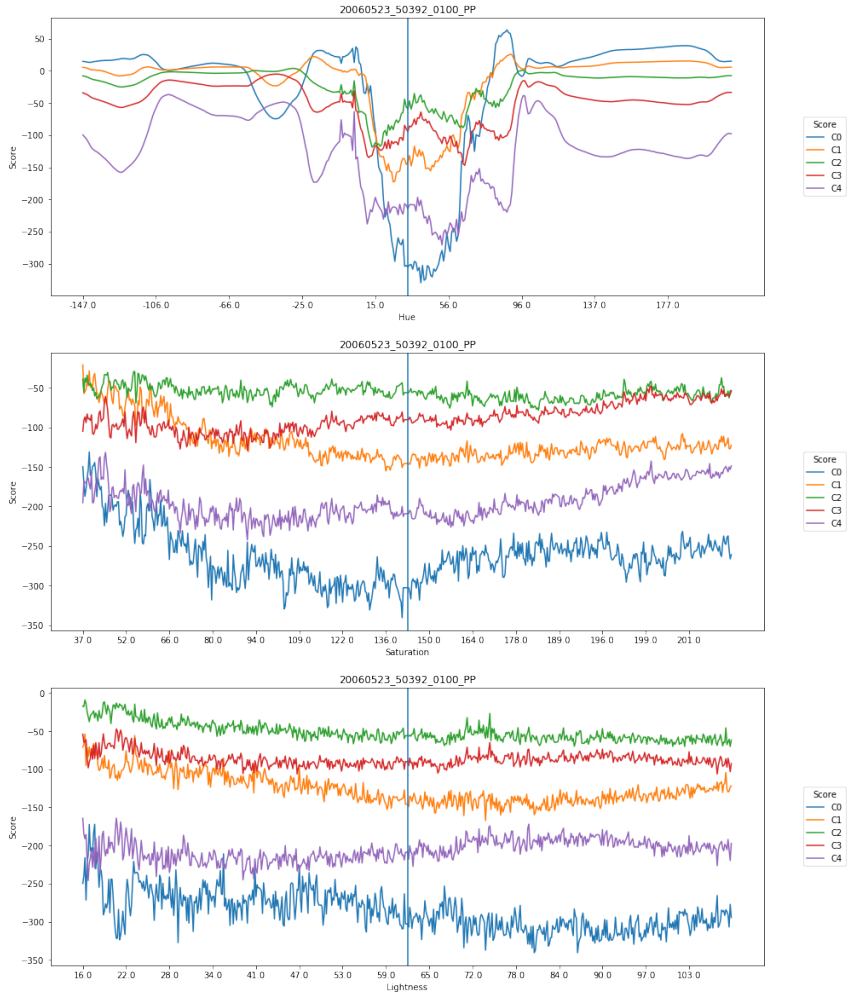


Hue — Saturation — Lightness

Stability analysis sample

20060523 50392 0100 PP Target: 2 HSL

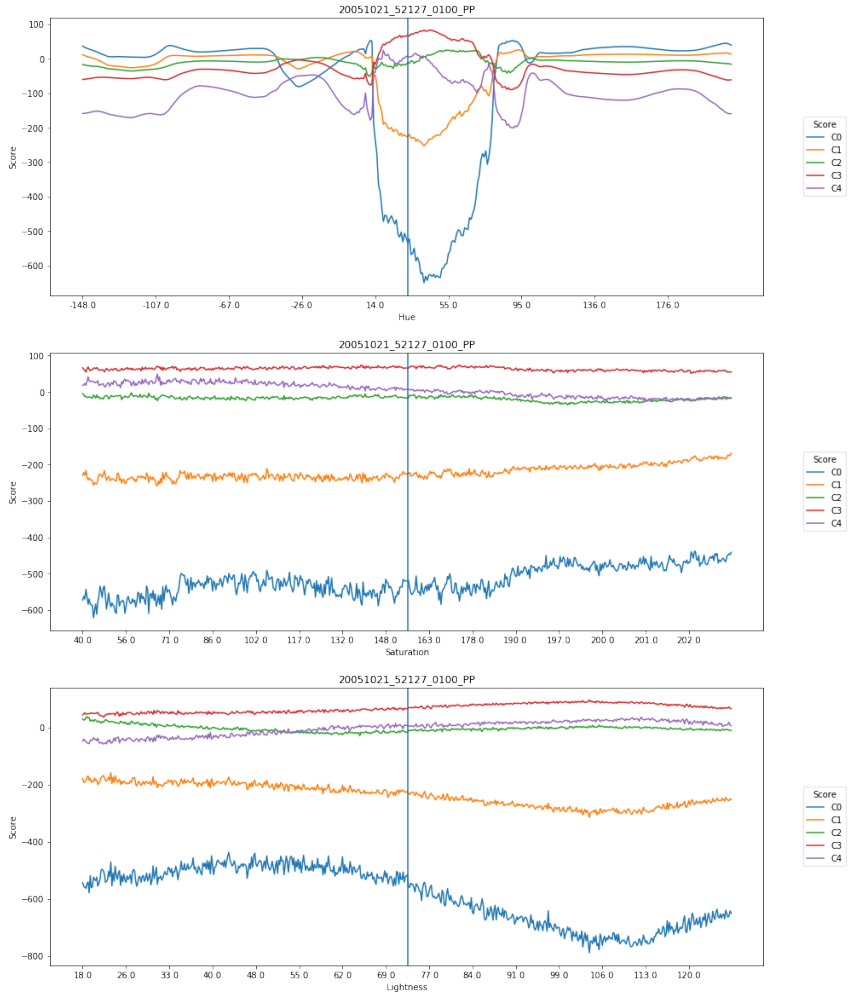




Hue — Saturation — Lightness

Stability analysis sample

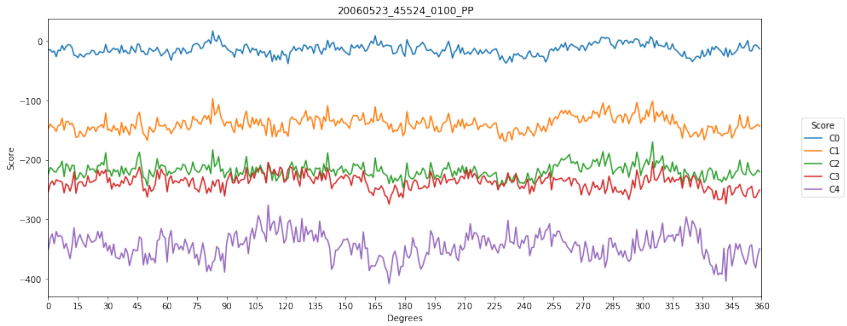
20051021 52127 0100 PP Target: 3 HSL



Hue — Saturation — Lightness

Stability analysis sample

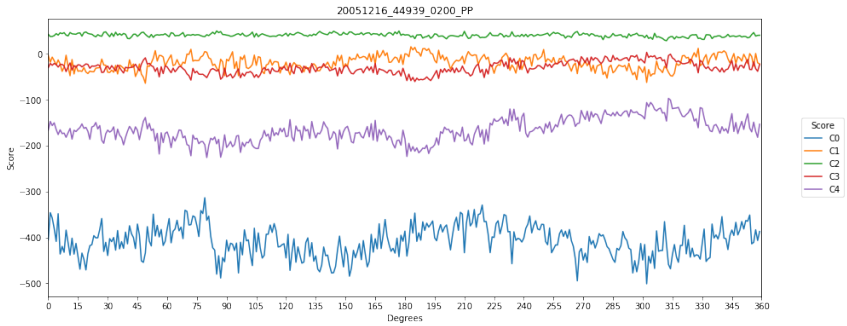
20060523 45524 0100 PP Target: 0 Rotation



Rotation Visualization

Stability analysis sample

20051216 44939 0200 PP Target: 2 Rotation



Rotation Visualization