NMF for quality control of multi-modal retinal images for diagnosis of diabetes mellitus and diabetic retinopathy*

Anass Benali[1,2], Laura Carrera[1,2], Ann Christin[1,2], Ruben Martín[1,2,6], Anibal Alé[3], Marina Barraso[3], Carolina Bernal[3], Sara Marín[3], Silvia Feu[3], Josep Rosinés[3], Teresa Hernandez[3,6], Irene Vilá[3,6], Cristian Oliva[3,6], Irene Vinagre[4,5,6], Emilio Ortega[4,5,6], Marga Gimenez[4,5,6], Enric Esmatjes[4,5,6], Javier Zarranz-Ventura[3,4,6], Enrique Romero[1,2], and Alfredo Vellido[1,2]

[1] Intelligent Data Science and Artificial Intelligence (IDEAI-UPC) Research Center
[2] Computer Science Department, Facultat d'Informàtica de Barcelona (FIB), Universitat Politècnica de Catalunya (UPC BarcelonaTech)
[3] Institut Clínic d'Oftalmología (ICOF), Hospital Clínic de Barcelona, Barcelona, Spain
[4] Diabetes Unit, Hospital Clínic de Barcelona, Barcelona, Spain
[5] Institut Clínic de Malalties Digestives i Metaboliques (ICMDM), Hospital Clínic de Barcelona, Barcelona, Spain
[6] August Pi i Sunyer Biomedical Research Institute (IDIBAPS), Barcelona, Spain

Abstract. In current ophthalmology, images of the vascular system in the human retina are used as exploratory proxies for pathologies affecting different organs. In this brief paper, we use multi-modal retinal images for assisting diagnostic decision making in diabetes mellitus and diabetic retinopathy. We report the use of matrix factorization-based source extraction techniques to pre-process the images as a data quality control step prior to their classification. Through this quality control, we unveil some relevant sources of bias in the data. After correcting for them, promising pathology classification results are still obtained, which merit further exploration.

Keywords: retinal imaging · non-negative matrix factorization · source extraction · medical data quality control · diabetes mellitus · diabetic retinopathy.

1 Introduction

Diabetic Retinopathy (DR) is the leading cause of human blindness in Type 1 Diabetes Mellitus (DM) patients. It is a serious and lifelong condition, estimated to constitute just between 5 to 10% of all diabetes cases [4] and characterized by the pancreas inability to generate enough insulin. When related complications affect the blood vessels in the retina, it can develop in what its known as DR,

* This research is partially funded by research grant PID2019-104551RB-I00.

Fig. 1. Fundus Retinography.

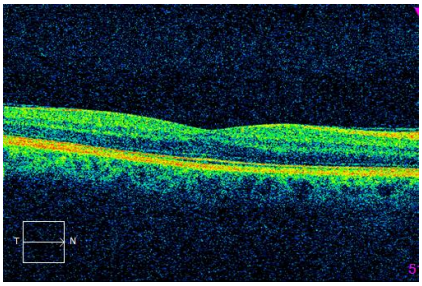

Fig. 2. Optical Coherence Tomography.

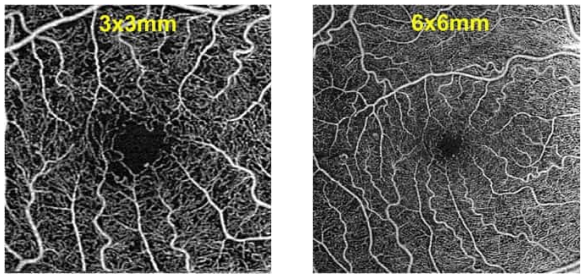

Fig. 3. Optical Coherence Tomography Angiography images at two resolution levels.

which may cause several vision difficulties [12]. In its early stages, DR may cause no symptoms, making it hard to detect and diagnose, as the differences between a healthy eye and an eye with early-stage DR are not obvious. Consequently, detecting DR early on after onset is key to slow its advancement, or even prevent the vision complications which can lead to blindness if left untreated.

Different non-invasive imaging techniques can be applied to the study of retinal diseases in general, and DR in particular, such as fundus retinography (FR, Figure 1), structural Optical Coherence Tomography (OCT, Figure 2), or Optical Coherence Tomography Angiography (OCTA, Figure 3).

Standard DR screening systems use FR [8] due to its widespread availability. For this reason, the vast majority of computational science applications in ophthalmology have been applied to FR images and, far more rarely, to OCT scans, which are not so readily available. Recently, the advent of the more advanced (and still scarcely used) OCTA technology allows for direct visualization of flow in the retinal vessels, easing the evaluation of these patients.

We analyze a high-quality multi-modal image dataset gathered in previous ophthalmology research projects [1–3, 14]. The ultimate goal of its analysis is diagnostic classification to be provided to the expert for decision-making assistance. In this study, though, we focus on image quality control prior to classification, using source extraction techniques, namely non-negative matrix factor-

ization (NMF) and some of its variants. This quality control is a necessary step to guarantee the robustness of the diagnostic assistance tool.

2 Materials and Methods

There are three image modalities in the analyzed dataset, namely FR, OCT and OCTA. For the OCTA images there are, in turn, four sub-modalities: $3 \times 3mm$ *superficial*, $3 \times 3mm$ *deep*, $6 \times 6mm$ *superficial* and $6 \times 6mm$ *deep*. Here *superficial* and *deep* refer to the perifoveal superficial capillary plexus (SCP) and perifoveal deep capillary plexus (DSP), respectively. This dataset is quite unique in that it only includes patients with Type 1 diabetes and that it also includes a subset of controls with no diabetes.

The DR scale is redefined to include such controls as *class 0* and, therefore, we have classes *0: Controls*, *1: No DR*, *2: Mild Non-proliferative DR*, *3: Moderate Non-proliferative DR*, *4: Severe Non-proliferative DR*, and *5: Proliferative DR*.

The original dataset includes 599 people in total. The retinal images acquired with the three imaging techniques (FR, OCT, OCTA) are available for both the left and right eye (whenever possible).

For a variety of reasons some of the images (and related clinical information) were missing in the original dataset. Therefore, the data needs some preliminary filtering. Also, some of the OCT, OCTA, FR scans are corrupted. Some examples of this can be seen in Figure 4.

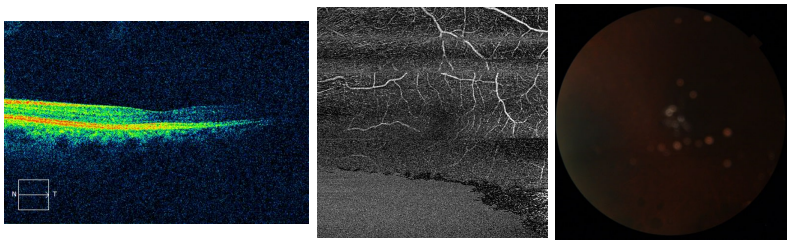

Fig. 4. Examples of corrupted OCT, OCTA and FR images.

Besides, some of the eyes on the dataset have other eye pathologies or previous treatments. To avoid biasing the models, only people with good enough quality OCT and OCTA were included. Also, patients that underwent treatment or surgery that can affect the captured features were filtered out.

The pre-filtering process for the OCT and OCTA was performed as specified in a previous study [2]. The FR images do not have any quality information and so they are all included in this filter.

After applying the exclusion criteria, as graphically described in Figure 5, we are left with 771 eyes.

Inspecting the distribution of the instances in the dataset (See Figure 6) and their distribution of exclusion (See Figure 7), too few class 3, class 4 and class

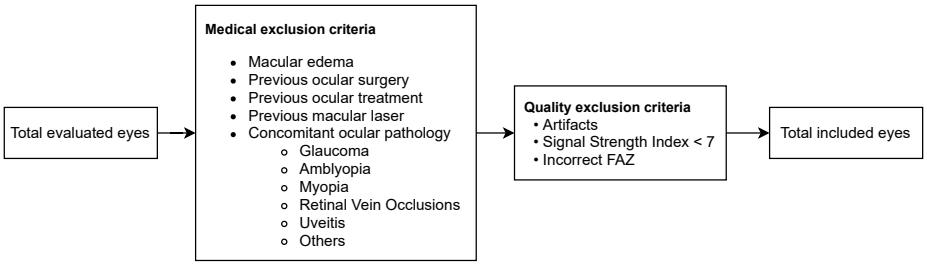

Fig. 5. Diagram showing the exclusion criteria [2].

DR scale	No. eyes before filter	No. eyes after filter
0	228	136
1	610	445
2	245	162
3	42	25
4	5	2
5	44	1

Fig. 6. DR class counts.

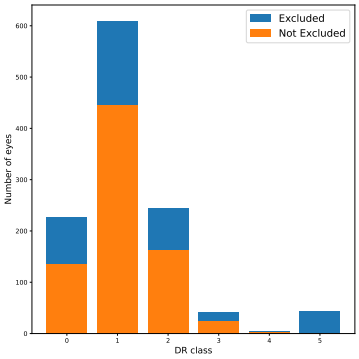

Fig. 7. Distribution of exclusion by class labels.

5 eyes are found to remain after filtering. Thus, from a classification viewpoint, it makes more sense to aggregate them into three classes: class 0, class 1 and class {2,3,4,5}, representing Controls, Type I diabetic with no DR and Type I diabetic with DR, respectively.

The ultimate goal of this investigation is the assessment of the capabilities of discrimination in the following two binary classification problems:

- **classification task 0-15**: Discrimination between the eyes of non-diabetics (controls) and of Type 1 diabetics, which corresponds to class {0} *versus* class {1,2,3,4,5}.
- **classification task 1-25**: Discrimination between no DR eyes and clinical DR eyes, corresponding to class {1} *versus* class {2,3,4,5}.

We will start by building small standard NMF models for a first exploration of the quality of the extracted image sources, removing those which are artefactual and reconstructing the target image through a linear combination of the relevant ones. In this setting, we prefer sources to capture sparse localized features so that the parts-based representation is easier to interpret for medical experts.

After this pre-processing, several models of the NMF family, namely standard NMF [10], Sparse NMF [9], Separable NMF [6] and Convex NMF [5] will be built

for each of the six types of image. The rank r (number of extracted sources) will be estimated by inspecting the decay of the SVD eigenvalues. We will retain a relatively large number of components so as not to restrict too much the features the models can learn.

For each combination of image type, model and parametrization, the data matrix decomposition obtained consists of sources and weights (encoding). Following similar approaches [13] that used NMF as the basis for subsequent classification, the encoding matrices of all NMF variants will be used as input variables in preliminary feature selection and classification models.

3 Exploratory Pre-processing

The first exploratory step consists on experimenting with small standard NMF models for each type of image and inspecting the extracted sources. The result of this step reveal some issues that should be fixed before using the encoding matrices of NMF for classification.

3.1 Retinography

Some small standard NMF models were built from the FR photographs, where images were initially modeled in a RGB (Red, Green, Blue) colour space. From Figure 8, we see that the extracted sources only separate the colour channels, the shades and illuminations, instead of intrinsic anatomical differences. Therefore, we will need to normalize the illumination in order to learn more effective features. This can be achieved with local adaptive filters. Here, we make use of the FR image pre-processing method made public by the winner of the Kaggle 2015 FR images competition [7].

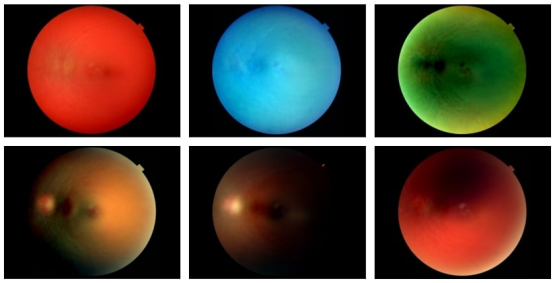

Fig. 8. Six sources of the FR exploratory model

3.2 OCT

Looking at the sorces reported in Figure 9, it seems clear that the NMF model is actually learning different translations and rotations of the OCT scans, which,

again, is not what we sought. Moreover, the strange-looking last source is actually the result of some OCT scans being in a grayscale colour space instead of RGB. An expert ophthalmologist confirmed that OCT scans were originally grayscale and that the colour in the image is extraneous. Therefore, they were all transformed to grayscale.

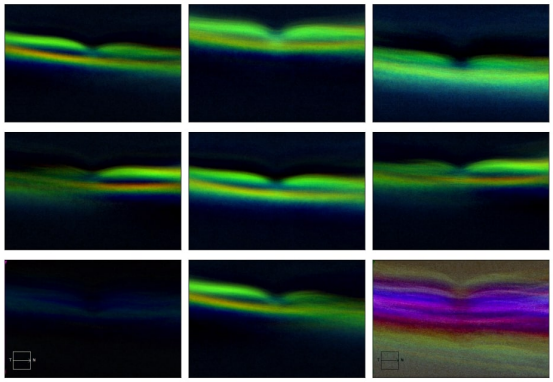

Fig. 9. Nine OCT sources of the exploratory model

To learn more intrinsic features instead, we developed a pre-processing step to isolate the regions of interest (ROI). The existence of a legend and symbol on the bottom-left of the OCT scan was also noticed. This recurrent feature was being isolated in its own source by NMF. Interestingly, the artefactual patterns present in the OCT scans (a magenta bar and the legend) where not identified as completely individual sources until setting the models to extract 13 sources (rank $r = 13$).

3.3 OCTA

OCTA images were found to be mostly fine to be used as they are. In Figure 10, some of the extracted sources are shown. Nevertheless, we still used noise reduction filters such as a median filter, bilateral filter and different types of image thresholding to improve the images.

The model still identified artefacts in some of the images. For example, one source helped identifying six images (see Figure 11) containing the camera model watermark on the bottom right. It also identified the eye of a patient with a very unusual path of the eye nerve through the FAZ area, as seen in Figure 12. All these edge cases were fixed or filtered out.

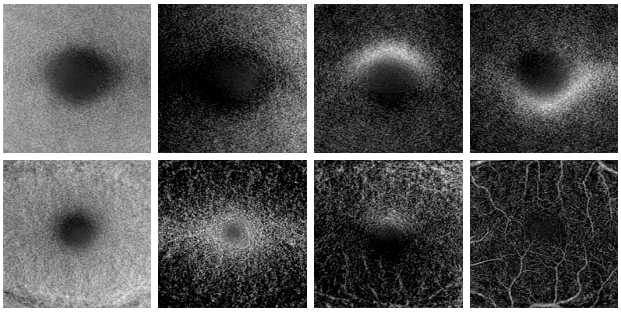

Fig. 10. Top row: OCTA *deep* sources. Bottom row: OCTA *superficial* sources.

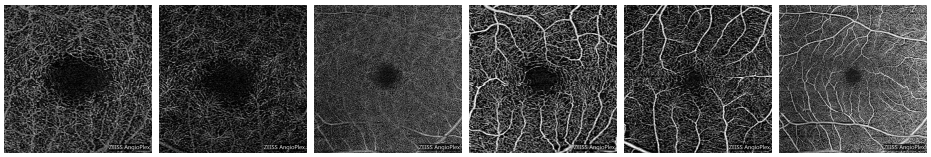

Fig. 11. OCTA images with a watermark on the bottom right corner.

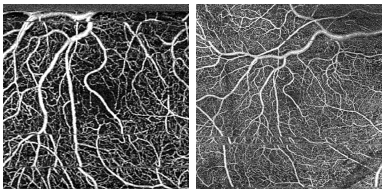

Fig. 12. Eye with unusual vessel pathing. Left: $3 \times 3mm$ Right: $6 \times 6mm$

4 Learning unsupervised part-based representations

Once the pre-processing of the images was carried out, a definitive extraction was implemented with all the models. Each NMF variant was run with different initializations (random, NNDSVD and NNDSVDa).

Regarding the number of components or sources to extract in the factorization models, since the ultimate goal is using the transformed data as the basis for a classification problem, a cross-validation (CV) scheme could be used for its choice. However, it would have to be tuned for each model and initialization. This is a rather cumbersome procedure and, therefore, we opted for a different strategy: a sensible range of values was found by inspecting the decay of the SVD eigenvalues (tantamount to looking at the retained variance of PCA) and was run for different numbers of sources. Supervised feature selection was then applied and used to decide which decomposition was best in terms of classification.

We tried to use the images at full resolution, but, for computational expediency, and given that preliminary results were not significantly different, image

sizes were reduced (for FR, to 256×256; for OCT, to 100×500; and for OCTA, to 256×256). With these, the SVD explained variance was calculated (see Figure 13).

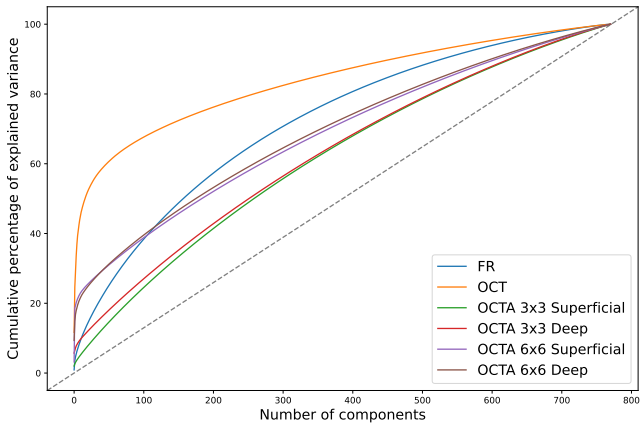

Fig. 13. SVD cumulative percentage of variance explained, for all image modalities.

It is noteworthy that OCT images show a "better" curve than the rest, as a low number of components already explain much variance. A possible reason for that could be that the factorization method does not agree well with the vessel variability present on the FR and OCTA images. Based on the plot, we decided to run the models for $r \in \{64, 128, 256, 384\}$. Going beyond that for most models would result in learning specific cases, instead of features. For the sparse models though, it could make sense to have a larger number of features, but we do not expect to have a large amount of relevant sparse and localized features in the data. The resulting encoding matrices were subsequently used as input features in classification designs. Then a feature selection and double CV scheme was performed to train and test the models.

4.1 Feature selection and classifier training

A supervised feature selection approach based on mutual information (MI) was applied and a stratified double-cross CV scheme was used to robustly train and test the classification models. Feature selection was carried for each of the classification tasks (diabetic or non-diabetic and presence of DR) and for each subset of features (FR, OCT, OCTA and all of them). The following ML/statistical classifiers were used: LR, LDA, Linear SVM and RBF-SVM.

Each selected subset of features was ordered from highest to lowest MI with the target class and the first 32 features were selected. This procedure uses a 10-fold CV; thus, the MI of each variable is computed for each CV split, resulting on 10 MI estimates for each variable that are then averaged.

Then, using a double CV scheme, the hyper-parameters of the classification models were optimized and the generalization error estimated. Specifically, the inner CV was used to select the best hyper-parameters according to the averaged validation AUC metric. Once the parameters of the models have been defined, a (optional) backward elimination wrapper method was applied to remove the irrelevant or less useful features for the model. In order to check if a feature can be safely removed, the hyper-parameters are re-optimized on the same corresponding inner CV to see if there is a decrease on the averaged AUC metric. Once the features and hyper-parameters are selected, a model is re-trained for each inner CV train split and are tested on the corresponding outer test CV fold.

The stratified double CV is defined with 5 splits on the outer CV and 4 splits of the inner CV (a total of 20 iterations). Using this scheme, 20 test estimates are obtained, which are displayed in the study as a boxplot. The grid search for the classifier hyper-parameters is shown in Table 1.

Method	Hyper-parameters
Logistic Regression	$C = 10^{-3:3}$
LDA	None
Linear SVM	$C = 10^{1:4}$
RBF-SVM	$C = 10^{1:4}$, $\gamma = 10^{-4:1}$

Table 1. Grid search values for the hyper-parameters of the classification methods. The notation $x : y$ denotes all the integers in the range $[x, y]$.

5 Results and Discussion

5.1 Useful sources learnt

After executing the models, we take a look at the learnt components. We show some of the relevant learnt standard NMF sources for each type of image when initialized with NNDSVD. Here, the NMF variants components are not shown because they are less intuitive to interpret and not always parts-based representations.

Some of the learnt NMF components for the FR images are represented in Figure 14. The sources mostly seem to capture the thickest vessels. Like in the OCTA sources, there is variability on the positioning of the vessels which ends captured in different sources.

Figure 15 shows some of the sources learnt by NMF for OCT images. We can see that the sources are a localized parts-based representation.

Looking at the learnt NMF components for the *deep* OCTA images (Fig. 16) reveals that they capture the different patterns around the FAZ area.

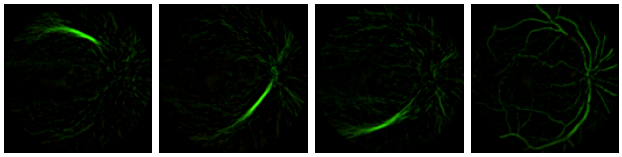

Fig. 14. Some NMF sources from FR images when initialized with NNDSVD

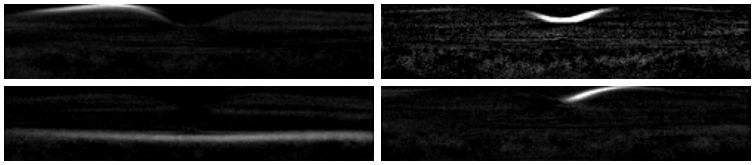

Fig. 15. Some NMF sources from OCT retinal images (initialized with NNDSVD)

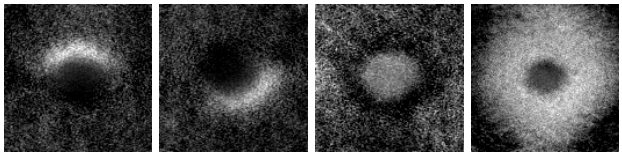

Fig. 16. Some NMF sources from OCTA *deep* images (initialized with NNDSVD)

The learnt NMF features for the *superficial* OCTA images can bee seen in Figure 17. A sparse representation is learnt. We notice that the bottom vessel is being captured by different components depending of its position.

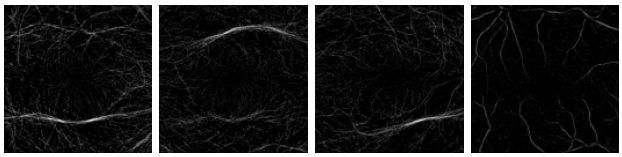

Fig. 17. Some NMF sources from OCTA *superficial* (initialized with NNDSVD)

5.2 Classification results

For comparison, a *Dummy Classifier* that generates predictions by following the training set class distribution is included. It will have an average AUC of 0.5.

We note that if the pre-processing explained on section 3 is not applied, the classification results are no better than random.

Discriminating DR For classification task 1-25 (see section 2), we obtain the results shown in Figure 18. OCT features yield the best results. The FR and OCTA features produce more or less similar results. The best results are obtained for logistic regression and LDA. We hypothesize that the reason SVM classifiers work worse is because the hyper-parameter search was not exhaustive enough, but this should be further investigated.

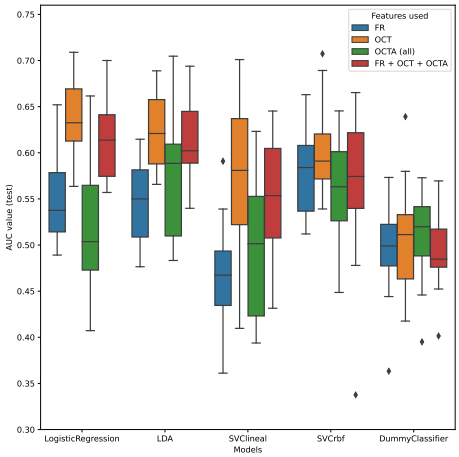

Fig. 18. Resulting boxplot of AUC results for classification *task 1-25*.

Discriminating DM For classification task 0-15, we obtain the results shown on the left plot in Figure 19. It can again be seen that the OCT features yield the best performance. The FR features work better than random, while the OCTA features are all over the place. At this point, we inspected the sources yielding the best results and found a bias in the data: that the range of images from 388 to 420 have OCT scans with an unexpected noise and level of gray. Some examples can be observed in Figure 22 and Figure 23.

This, in itself, would not be necessarily a problem if it was not because that range of images has more controls than the other classes. In the filtered data, those are 32 individuals of class 0, 4 of class 1 and 1 of class 2.

According to the ophthalmologist, a possible explanation for this could be that the lens of the camera equipment was dirty when those images were taken, or an artefact in the export from the camera equipment software. We decided to test how the model performs when removing those instances. This change means reducing the number of *class 0* eyes from 136 to 104. By doing so, the results shown on the right hand side plot of Figure 20 are obtained. They worsen slightly for OCT and have higher variance. Oddly enough, the results for retinography marginally improved. Since there is no quality filter for the FR images, it could

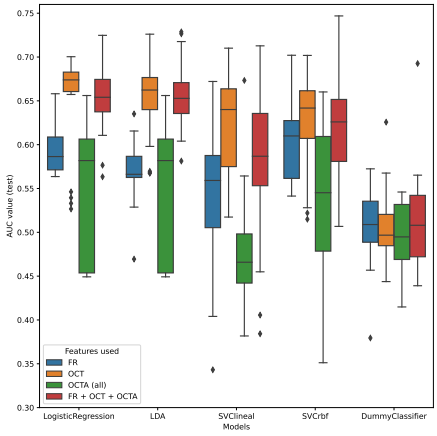

Fig. 19. Boxplot of *task 0-15*

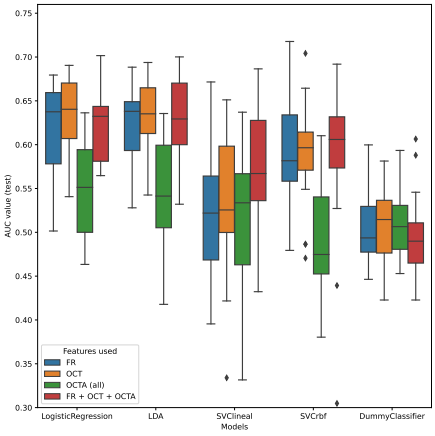

Fig. 20. Boxplot of *task 0-15* (without bias)

Fig. 21. Boxplots of the results when discriminating DM from controls

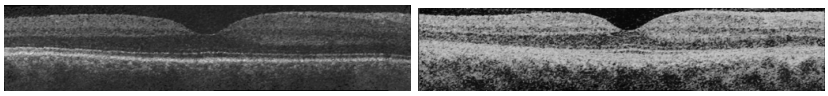

Fig. 22. Example of the found bias in the preprocessed data

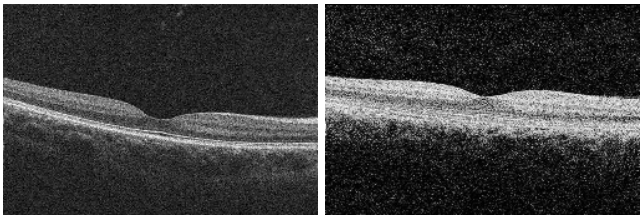

Fig. 23. Example of the found bias on the original images

be that the removed instances were difficult ones where the models previously failed. Also, we notice that, although the OCTA results worsen slightly on average, they exhibit lower variance.

Discriminating controls from DR For completeness, we perform the classification task class 0 versus {2,3,4,5}. Hence, we remove *class 1* (the majority class), emphasizing the importance of the detected bias of *class 0*. We call this classification *task 0-25*. Experiments were run with and without the biased data. The results are reported in Figures 24 and 25.

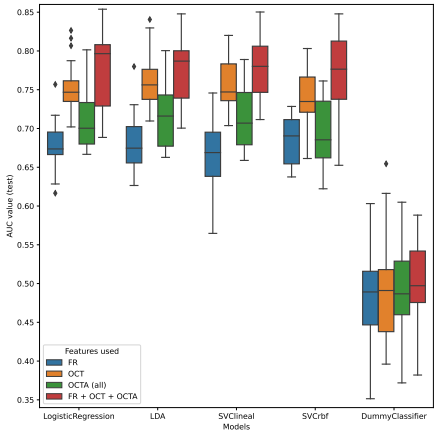

Fig. 24. AUC boxplot for *task 0-25*.

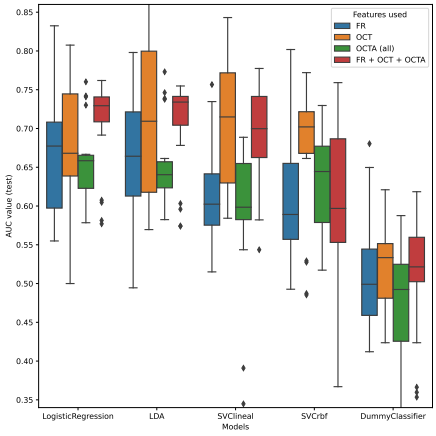

Fig. 25. AUC boxplot for *task 0-25* (without bias).

The results are consistently better and with less variability when including the biased image range: all the models perform similarly (even SVM, which yielded worse results in the other tasks). Removing the biased range decreases the performance and increases the variability of the results, but it is still far better than the dummy classifier. Worth stressing that the use of all the features (FR, OCT, OCTA) gives consistently the best results, with and without bias. This task showcases the importance of having found the bias on the image data and exemplifies the potential of NMF as a quality control tool.

6 Summary and Conclusion

In this brief paper, we have shown how a matrix decomposition method for feature extraction, namely NMF, can successfully be used for quality control in a medical imaging problem, by detecting data artefacts and biases. These methods can also provide further insight into the images themselves, increasing the interpretability of the results, a requisite for the application of ML models to medical problems [11].

The results reported in the previous section indicate that the NMF description of the image data is capable of discriminating in the tasks posed, albeit with varied results. OCT with logistic regression yields the best results in task 0-15 (between 0.6 and 0.65 average AUC) and, most interestingly, for the more difficult problem of task 1-25 (between 0.65 and 0.70), which opens the door to early DR warnings for patients already with Type I DM. Encouragingly, the discrimination between controls and DR patients consistently reaches average AUC values over 0.70 with the use of all modalities. Logistic regression is a tried and trusted model in the medical domain, which could provide an extra push to the

practical implementation of the NMF based analytical pipeline. With the experiments for task 0-25 we have shown how a significant bias found using NMF-based feature extraction affects the results. This exemplifies how important is quality control and how NMF can help to identify data issues given the right settings. Overall, these promising results warrant further investigation and comparison with alternative feature extraction methods and data representations.

References

1. Alé-Chilet, A., Bernal-Morales, C., Barraso, M., et al.: Optical coherence tomography angiography in type 1 diabetes mellitus. Report 2: Diabetic kidney disease. Journal of Clinical Medicine **11**(1), 197 (2022)
2. Barraso, M., Alé-Chilet, A., Hernández, T., et al.: Optical coherence tomography angiography in type 1 diabetes mellitus. Report 1: diabetic retinopathy. Translational Vision Science & Technology **9**(10), 34–34 (2020)
3. Bernal-Morales, C., Alé-Chilet, A., Martín-Pinardel, R., et al.: Optical coherence tomography angiography in type 1 diabetes mellitus. Report 4: Glycated haemoglobin. Diagnostics **11**(9), 1537 (2021)
4. Daneman, D.: Type 1 diabetes. The Lancet **367**(9513), 847–858 (2006)
5. Ding, C.H., Li, T., Jordan, M.I.: Convex and semi-nonnegative matrix factorizations. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(1), 45–55 (2008)
6. Gillis, N., Vavasis, S.A.: Fast and robust recursive algorithms for separable nonnegative matrix factorization. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(4), 698–714 (2013)
7. Graham, B.: Kaggle diabetic retinopathy detection competition report (2015), https://www.kaggle.com/c/diabetic-retinopathy-detection/discussion/15801
8. Grzybowski, A., Brona, P., Lim, G., Ruamviboonsuk, P., Tan, G.S., Abramoff, M., Ting, D.S.: Artificial intelligence for diabetic retinopathy screening: a review. Eye **34**(3), 451–460 (2020)
9. Kim, H., Park, H.: Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics **23**(12), 1495–1502 (2007)
10. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
11. Lisboa, P., Saralajew, S., Vellido, A., Villmann, T.: The coming of age of interpretable and explainable machine learning models. In: Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2021). pp. 547–556 (2021)
12. National Health Service UK: Diabetic retinopathy (2018), https://www.nhs.uk/conditions/diabetic-retinopathy/
13. Núñez, L.M., Romero, E., Julià-Sapé, M., Ledesma-Carbayo, M.J., Santos, A., Arús, C., Candiota, A.P., Vellido, A.: Unraveling response to temozolomide in preclinical gl261 glioblastoma with MRI/MRSI using radiomics and signal source extraction. Scientific Reports **10**(1), 1–13 (2020)
14. Zarranz-Ventura, J., Barraso, M., Alé-Chilet, A., et al.: Evaluation of microvascular changes in the perifoveal vascular network using optical coherence tomography angiography (octa) in type i diabetes mellitus: a large scale prospective trial. BMC Medical Imaging **19**(1), 1–6 (2019)