Medical Image Grading with Deep Quantum Ordinal Regression

Santiago Toledo-Cortés, Diego H. Useche, Henning Müller, and Fabio A. González

MindLab Research Group, Universidad Nacional de Colombia, Bogotá, Colombia
{stoledoc, diusecher, fagonzalezo}@unal.edu.co

Abstract. Although for many diseases there is a progressive diagnosis scale, automatic analysis of medical images is quite often addressed as a categorical or even binary classification problem, missing the finer distinction and intrinsic relation between the different possible stages or grades. Ordinal regression (or classification) considers the order of the values of the categorical label and thus takes into account the order of grading scales used to asses the severity of different medical conditions. This paper presents a probabilistic deep learning ordinal regression model for medical image classification that takes advantage of the representational power of deep learning and of the intrinsic ordinal information of disease stages by means of a differentiable probabilistic regression method. Approaching the problem as an ordinal regression task not only improves the final accuracy of the model, but also the interpretability of the results by means of a prediction uncertainty quantification, when compared to conventional deep classification and regression architectures.

Keywords: Deep Learning · Density Matrices · Diabetic Retinopathy · Eye Fundus Images · Histopathology Images · Ordinal Regression · Prostate Cancer · Quantum Measurement · Uncertainty Quantification

1 Introduction

Stages of a disease are not categorical. The progress of the degenerative process of a disease is not a jump from one class to another but the advance along a continuous route [24]. The possible stages of a disease are then the result of an effort by specialists to discretize a continuous behaviour. While not completely accurate, this information is of great utility in the generation of automatic systems if a model with the appropriate descriptive capability is used. The grade-based labelling of a disease contains information that is obviously discarded if a multiclass categorical classification model is used. The way to exploit the information of the grades of a disease is therefore through a regression modelling. If, in addition, a probabilistic regression model is used, the model predictions can be interpreted as probability distributions over the domain of the labels. With this, more general models can be achieved, with the capability to infer the disease

stage in a non-categorical way, and the capability to offer more information about the prediction by means of the uncertainty quantification.

Deep convolutional neural networks (CNN) represent the state of the art in the analysis of visual information [35]. Their advantage is based on the structure that gives them their name: convolutional filters, of which the parameters are learned and that make it possible to recognise geometric patterns in images. These patterns become increasingly complex as the model becomes deeper. In this way, we end up with models capable of recognising specific complicated objects, or in the case of medical images, capable of classifying an image according to a disease, or localizing a region of interest. On top of the convolutional blocks, a set of dense layers of neurons makes the model suitable for a classification task, using in the output layer as many neurons as there are categories, and cross-entropy as loss function. For a regression task, a single neuron output is needed, using a loss function derived from the absolute error.

In the medical field, deep CNN's have been shown to be effective for analysing images and visual content of all kinds. From X-rays to diagnose osteoporosis, to MRI's to diagnose brain diseases. Although many diseases present different stages on a progressive scale and in many cases this information is available, a binary labelling is usually favoured [24]. While binarizing these diagnosis tasks may have sense from a treatment-oriented decision, two drawbacks arise: first, the ordinal information of the grades is not taken into account for the training process. Second, the model predictions, usually subject to a softmax activation function in a neural network model, cannot be interpreted as a probability distribution [42].

In this paper we present the Deep Quantum Ordinal Regressor (DQOR), a deep probabilistic model that combines a CNN with a differentiable probabilistic regression model, the Quantum Measurement Regression (QMR) [12]. This approach allows us to:

1. Predict posterior probability distributions over the grades range. Unlike other probabilistic methods such as Gaussian processes, these are explicit discrete distributions.
2. In the case of patch-based analysis, integrate patch posterior distributions into a single whole-slide image distribution in a simple, yet powerful probability-based manner.
3. Quantify the uncertainty of the predictions. This enriches the model as a diagnostic support tool, which in safety-critical applications, provides the method with a first level of interpretability.
4. Improve a prognosis-oriented binary diagnosis, based on an ordinal-label end-to-end training.

To show the effectiveness of our proposal, we test it on two stage-based diagnosis tasks: prostate cancer (PCa) diagnosis, and Diabetic retinopathy (DR) diagnosis. PCa is currently the second most common cancer among men in America. Early detection allows for greater treatment options and a higher chance of treatment success, but while there are several methods of initial screening, a

concrete diagnosis of PCa can only be made with a prostate biopsy [9]. Tissue samples are currently recorded in high-resolution images, called whole-slide images (WSIs). In these images the pathologists analyze the alterations in the stroma and glandular units and identify the present Gleason patterns in a scale from 1 to 5. The sum of the two most dominant Gleason patterns gives the final Gleason score. Then the possible range for this Gleason score is from 2 to 10. However in practice the specialists only care about the highest five grades, from 6 to 10, as for low grades below 3 biopsies are not taken [22]. The higher the grade, the more advanced the cancer. Although the CNN implementation in automatic classification models for PCa has been widely studied, there is still much research to be done in relation to the diagnostic process in histopathology [37], and the usual approach is as a multi-class or a binary classification of low risk (6-7 GS) vs high risk (8-10 GS) cases [19].

Something similar happens with DR, which is a consequence of diabetes mellitus (DM), one of the most prevalent diseases worldwide [23]. Early DR diagnosis allows to prevent most of the severe consequences of the disease, including complete blindness [46]. One method for early and effective diagnosis of DM consist of the inspection of the retinal tissue. This is made by means of an eye fundus image, an RGB photo of the inner back of the eyeball that allows to detect specific lesions that are a direct consequence of the alterations caused by diabetes. These lesions include microaneurysms, neovascularization, hemorrhages, exudates, etc. From the count of these lesions, an ophthalmologist specialist in retina can give a diagnosis of the disease, on a five-level severity scale from zero to four (0→4), being zero (0) a negative case of DR, and four (4) a case of proliferative DR (see Figure 1) [3]. Many approaches have treated the problem from a multi-class classification perspective, or simple as a binary task, where a diagnosis of 0 or 1 grade corresponds to a case of *non-referable* DR and a case of 2, 3 or 4 grade corresponds to a case of *referable* DR [36].

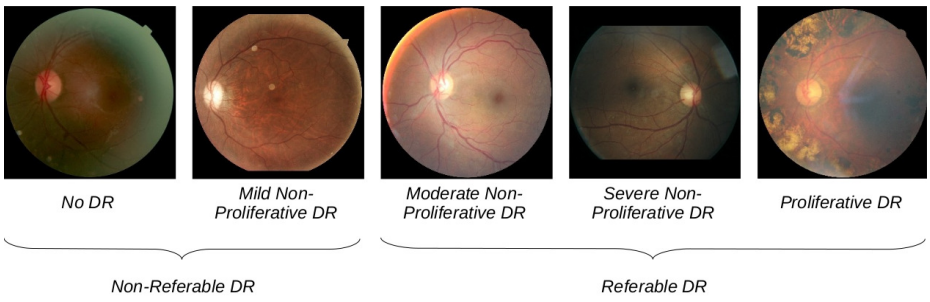

Fig. 1. Five possible grades for DR diagnosis. No DR cases correspond to grade 0. Grades 0 and 1 correspond to *non-referable* DR cases, while grades 2, 3 and 4 correspond to *referable* DR. Samples extracted from EyePACS dataset [8].

The paper is organized as follows: Section 2 presents a brief overview of the related work. Section 3 presents the theoretical framework of the DQOR, and Section 4 presents the experimental set up. In order to validate our approach, we compare our performance with state of the art deep learning-based methods [19] [40], and with closely related classification and regression methods. Section 5 presents the experimental results and finally in Section 6 we present the conclusions of this work.

2 Related Work

Ordinal regression tasks are not exclusive to the medical field. Therefore the development of ideas for dealing with this kind of label structure has occurred alongside the development of the rest of machine learning algorithms, as an intermediate field between regression and classification models. According to Gutierrez et al. [14], a taxonomy of the methods previously proposed for this goal can be made as follows: first, *naïve* approaches, which are basically standard machine learning models for nominal classification or metric regression. Second, *ordinal binary decomposition* approaches, which break down the problem into several binary sub-problems [10]. And third, in which our proposal is framed, *threshold* models, which are based on a predictor that yields a real value, which is then approximated to an integer value. Depending on the particular problem, different models may perform better than others, so there is not a best overall ordinal regression approach [14].

In the medical field, while it is true that there have been some applications of ordinal regression models, there is no clear and well-defined trend. Recently, for instance, the ordinal regression by binary classifiers has been applied to Facial Age Estimation [28] [38], and diagnosis of Alzheimer's disease [21], taking advantage of the inherent ordinal severity of brain degeneration. However, in addition to the value of prediction, something closer to a real diagnosis given by a specialist, is to be able to obtain a concrete probability distribution over the possible stages of the disease. And as a matter of fact, the distribution describing the probability of belonging to a disease stage has to be unimodal, so it makes sense to use unimodal distributions in ordinal classification tasks [6]. Models have already been used where predictions are forced to follow a Poisson or binomial distributions over the possible outputs [6], showing that, when needed, the ordinal approach improves results when compared to a cross-entropy approach. Beckham and Pal used the same idea in [5], where the last layer of a neural network is interpreted or forced to behave like a Gaussian distribution, and the final prediction value is the expected value of the neural network.

Regarding our particular interest of application, most of the recent work for PCa is focused on classifying Whole Slide Images (WSI) by low and high GS [19]. To train a model on WSIs, it is required to divide each image into multiple patches and then to summarize the information of the patches by different methods, hence obtaining a prediction of the WSI. In [15], the authors clas-

sify patches between low, and high GS, using various CNNs, and summarizing the patches to a WSI by a GS majority vote. Another approach by Tolkach et al. [41] uses a NASNetLarge CNN, and summarizes the GS of the patches by counting the probabilities per class. In Karimi et al. [16] they proposed training three CNNs for patches of different sizes, and summarizing the probabilities by a logistic regression.

Some other CNN architectures for GS grading include a combination of an atrous spatial pyramid pooling and a regular CNN as in [22], an Inception-v3 CNN with a support vector machine (SVM) as in [25] and a DeepLabV3+ with a MobileNet as the backbone [18]. In [27], the authors use an InceptionV3 with a k-nearest-neighbor classifier to summarize the heatmap. Other techniques for GS grading include, Support Vector Machine Feature-Recursive Feature Elimination [33], and learning features from *bag-of-words* features [45].

On the DR side, much work has focused on a binary diagnosis based on deep neural networks [29]. Toledo-Cortés et al. [40] used the model proposed in [43] to extend a neural network-based binary classifier into a grading regressor by means of a Gaussian process. Tian et al. [39] also use a deep CNN as backbone for a model trained to optimize a combination of a metric loss and a focal loss function for soft labels, in an attempt to use the ordinal information of the DR stages. For their part, Teresa Araujo et al. [4] propose DR|GRADUATE, a deep learning based model in which the last layer has as many neurons as classes, a Gaussian filter is performed on that output, and the model is trained with a loss function that controls the entropy of the classification on the one hand, and the standard deviation of the distribution on the other. This allows not only to infer the DR grade, but also to measure the uncertainty associated with the prediction.

Related to the ordinal regression, uncertainty quantification was analysed in many studies with the aim of obtaining more interpretable models in circumstances where it is necessary to gain the trust of the end user [34]. Studies were conducted on the uncertainty of machine learning algorithms for organ classification [26] and estimation of tissue parameters in the operating room [2]. Also, estimating the uncertainty of a model's prediction is of interest as it helps to avoid the consequences of blind use of the model's inference [17]. This is particularly true in medical settings, where misdiagnosis can have serious consequences for patients. With this in mind, Leibig et al. [20] analysed uncertainty information from deep neural networks for DR detection. The authors tested dropout-based Bayesian uncertainty estimation against alternative techniques, such as direct analysis of the softmax output of the network, claiming that Bayesian approaches perform better for uncertainty estimation and also showing that uncertainty-aware decisions can improve the overall grading process.

3 Deep Quantum Ordinal Regressor

The overall architecture of the proposed Deep Quantum Ordinal Regressor (DQOR) is described in Figure 2. We use a deep CNN as a feature extractor. The extracted

features are then used as inputs for the QMR method [12]. QMR uses density matrices for regression problems and works as a non-parametric density estimator. It requires an additional feature mapping from the inputs to get a quantum state-like representation. This is made by means of a random Fourier features approach [30]. The regressor yields a discrete posterior probability distribution from which we get the final grade prediction and an uncertainty measure.

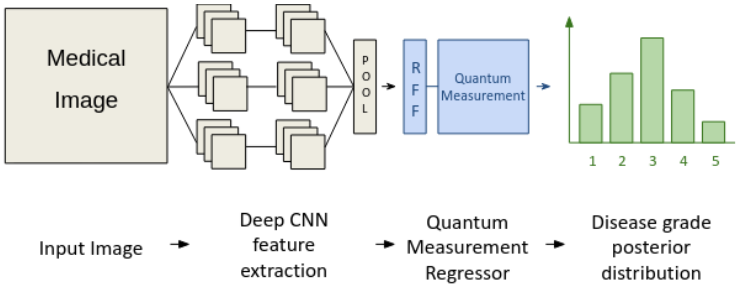

Fig. 2. Overview of the proposed DQOR method for medical image analysis. A deep CNN is used as feature extractor for the input image. Those features are the input for the QMR regressor model, which yields a posterior probability distribution over the possible grades of the disease.

3.1 Feature Extraction

Automatic image analysis regardless of the source, medical or not, relies on deep convolutional networks. The representational power of these models has allowed remarkable advances in computer vision and therefore we will use them as a basis for feature extraction [13]. Regardless of the used CNN, there is a basic structure that is always maintained: an input layer of the image, followed by a series of convolutional blocks and a pooling layer that summarises all the information extracted by the convolutions. Usually, this layer is followed by a series of dense layers that take care of the final classification of the image. We take the output of the pooling layer as a representation of the image.

3.2 Random Fourier Features

The random Fourier features (RFF) method [30] creates a feature map of the data $\mathbf{z}(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^{\mathbb{D}}$ in which the dot product of the samples in the $\mathbb{R}^D$ space approximates a shift invariant kernel $k(\mathbf{x} - \mathbf{y})$. The method works by sampling i.i.d. $w_1, \cdots, w_D \in \mathbb{R}^n$ from a probability distribution $p(w)$ given by the Fourier transform of $k(\mathbf{x} - \mathbf{y})$, and sampling i.i.d. $b_1, \cdots, b_D \in \mathbb{R}$ from a uniform distribution in $[0, 2\pi]$. In our context, the shift invariant kernel is the

Radial Basis Function (RBF) given by, $\mathrm{k}_{\mathrm{RBF}}(\mathbf{x}-\mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|^2}$, where gamma γ and the number D of RFF components are hyper-parameters of the models. In our model the RFF works as an embedding layer that maps the features from the deep CNN module to a representation space that is suitable for the quantum measurement regression layer.

3.3 Quantum Measurement Regression (QMR)

QMR [12] is a differentiable probabilistic regression model that uses a learnable density matrix, ρ_{train}, to represent the joint probability distribution of inputs and labels. A QMR layer receives a RFF encoded input sample $|\psi_x\rangle$, and then builds a prediction operator $\pi = |\psi_x\rangle\langle\psi_x| \otimes \mathrm{Id}_{\mathcal{H}_\mathcal{Y}}$ where $\mathrm{Id}_{\mathcal{H}_\mathcal{Y}}$ is the identity operator in $\mathcal{H}_\mathcal{Y}$, the representation space of the labels. Inference is made by performing a measurement on the training density matrix ρ_{train}:

$$\rho = \frac{\pi\rho_{\mathrm{train}}\pi}{\mathrm{Tr}[\pi\rho_{\mathrm{train}}\pi]}. \tag{1}$$

Then a partial trace $\rho_\mathcal{Y} = \mathrm{Tr}_\mathcal{X}[\rho]$ is calculated, which encodes in $\rho_{\mathcal{Y}rr}$, with $r \in \{0, \ldots, N-1\}$, the posterior probability over the labels. The expected value represents the final prediction $\hat{y} = \sum_{r=0}^{N-1} r\rho_{\mathcal{Y}rr}$.

A gradient-based optimization is allowed by a spectral decomposition of the density matrix, $\rho_{\mathrm{train}} = V^\dagger \Lambda V$, in which the number of eigen-components of the factorization is a hyper-parameter of the model. The model is trained by minimizing a mean-squared-error loss function with a variance term whose relative importance is controlled by hyper-parameter α:

$$L = \sum (y - \hat{y})^2 + \alpha \sum_r \rho_{\mathcal{Y}rr}(\hat{y} - r)^2. \tag{2}$$

3.4 Patch-based Analysis Summarization

Patch-based image analysis is preferred or rather needed for some applications. This is the case for the WSI analysis used for the prostatic cancer diagnosis. This implies an additional stage, as a summarizing is required to reach a whole image prediction. The simplest procedure would be a majority vote (MV), as reported in most of previous works. In the majority vote, the prediction for an image is decided according to the grade with the highest number of predictions among the patches of the image. DQOR allows, however, a probability vote procedure (PV): since each patch can be associated with a probability distribution, the normalized summation yields a distribution for the whole image. More formally, thanks to the law of total probability, given an image I, composed by n patches, each patch denoted by p_i, the posterior probability of the grade r is,

$$P(r|I) = \frac{P(r,I)}{P(I)} = \frac{\sum_{i=1}^{n} P(r|p_i,I)P(p_i|I)P(I)}{P(I)} = \frac{1}{n}\sum_{i=1}^{n} P(r|p_i). \tag{3}$$

The final prediction value may corresponds to the grade with highest probability or to the expected value of the distribution.

4 Experimental Set Up

The specific details of the experimental procedure carried out for each of the tasks are described below.

4.1 Prostate Cancer

The specific set up of the DQOR applied for Prostate cancer image analysis is described in Figure 3.

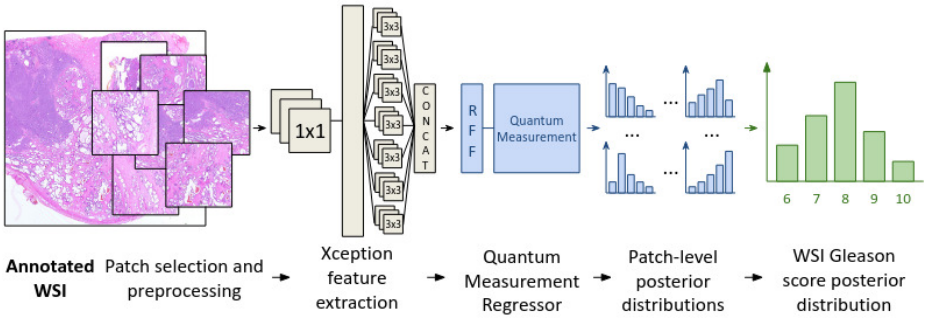

Fig. 3. Overview of the proposed DQOR method for prostate tissue grading. A Xception network is used as feature extractor for the image patches. Those features are the input for the QMR regressor model that yields a posterior probability distribution by patch over the Gleason scores. Then, those distributions are summarized into a single discrete probability distribution for the WSI.

Dataset We use images from the TCGA-PRAD dataset, which contains samples of prostate tissue with GS from 6 to 10. This data set is publicly available via The Cancer Genome Atlas (TCGA) [15]. In order to directly compare our results with our baseline [19] we use the same subset and partition for train and test, which is detailed in Table 1. The patch extraction details are described in [15].

The training of the model is made at the patch level. In order to obtain predictions at the WSI level, a process of summarization is carried out. Therefore, each patch is assigned the GS of the WSI of which it is a part. Although it makes no sense to say that a GS can be assigned to a single patch, the focus of our research is to show the effectiveness of the regression approach by comparing with previous work that handles labels in the same way, beyond specifically determining a Gleason pattern in each patch.

Feature Extraction The model presented in [19] is used as feature extractor. It is publicly available and consists of an Xception network trained on ImageNet

Table 1. Details of the subset and final partition of the TCGA dataset used for training and testing. This is the same partition used in [19].

Risk	Gleason Score	Train Samples	Validation Samples	Test Samples
Low	6	11	4	4
Low	7	53	17	17
High	8	23	8	8
High	9	50	17	16
High	10	4	2	1

and fine-tuned on prostate tissue image patches. This network was originally used for an automatic information fusion model for the automatic binary (low-high) classification of WSIs. The augmentation procedure and training details are described in [19]. Taking the output of the last average pooling layer of the model we got a 2048-dimensional vector representing each image patch.

Quantum Measurement Regression For the QMR, hyper-parameter tuning of the model was performed by generating 25 different random configurations. As result, we created an embedding with 1024 RFF components, 32 eigenvalues and γ was set to 2^{-13}. For the loss function (See eq. (2)), α was set at 0.4, and we set a learning rate of 6×10^{-5}.

Baseline An extension of the feature extractor model was set up as baseline for this work. Called *DLC-PCa* hereafter, it consists on 1024 neurons with ReLU as the activation function and a dropout of 0.2, followed by 5 neurons with a softmax activation function for the output. The learning rate was set to 10^{-7}, as in the baseline [19]. We also explored two closely related methods to QMR: Density Matrix Kernel Density Classification (DMKDC) [12] and Gaussian processes. DMKDC is a differentiable classification method, which applies a RFF feature map to the input sample, and then computes the expected value of the input with a density matrix of each class, returning a posterior probability distribution, which can be optimized with a categorical cross entropy loss function. All the previous experiments were performed in Python using the publicly available Keras-based implementation presented in [12].

We also explored a closely related method, a Gaussian process (GP) [31] which is another powerful Bayesian approach to regression problems. By means of a kernel covariance matrix, the GP calculates and iteratively updates the probability distribution of all the functions that fit the data, optimizing in the process the kernel parameters. In our case, we set the kernel as the RBF. The prediction process consists in marginalizing the learned Gaussian distribution of which the mean would be the actual prediction value and its standard deviation an uncertainty indicator. We performed experiments with GP using the Scikit-Learn implementation in Python. We also explored deep Gaussian processes (DGP), using the implementation proposed in [7], which also uses RFF

to approximate the covariance function. For those experiments, another hyper-parameter random search was made, finally setting the number of RFF to 1024 and the learning rate to 2×10^{-12} in a single layer schema.

4.2 Diabetic Retinopathy

The specific set up of the DQOR applied for eye fundus image analysis is described in Figure 4.

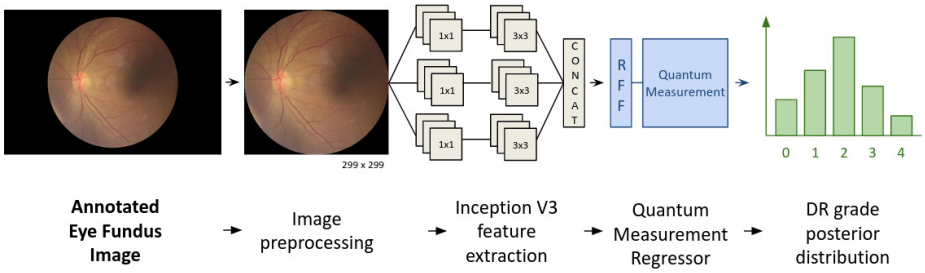

Fig. 4. Overview of the DQOR for diabetic retinopathy grading. An Inception-V3 network is used as feature extractor for the eye fundus image. These features are the input for the QMR regressor model, which yields a posterior probability distribution over the DR grades.

Datasets In order to directly compare the DQOR performance with state of the art baseline, we worked with EyePACS [8] and Messidor-2 [1] datasets. EyePACS is one of the largest publicly available datasets of eye fundus images. Each sample is labeled in a five grade scale from 0 to 4, where 0 stands for a healthy case, 1 for mild non-proliferative DR, 2 for moderate non-proliferative DR, 3 for severe non-proliferative DR and 4 for proliferative DR. In order to compare with our baseline [44] [40], we kept the same set up for training and test partitions, which are described in Table 2. Regarding Messidor 2, it has become an standard dataset in the field for testing. It consists of 1748 eye fundus images. While DR grades are not provided, we will use it to show the effectiveness of our proposal for a purely binary diagnosis. Details of Messidor-2 are described in Table 3.

Feature Extraction For the DR grading we use the model presented in [40], which is also publicly available, and consists of an Inception-V3 network trained on ImageNet and fine-tuned on eye fundus images from the EyePACS dataset. This network was originally used as feature extractor in a model for an automatic DR grading. In such a model, the training was made in two independent stages, one for the Inception-V3 and another for the Gaussian process. However, the

Table 2. Details of the subset and final partition of the EyePACS dataset used for training and testing. This is the same partition used in [40] and in [44].

Referable	Grade	Train Samples	Test Samples
No	0	37209	7407
No	1	3479	689
Yes	2	12873	0
Yes	3	2046	0
Yes	4	1220	694

Table 3. Details of Messidor-2 dataset used for testing. Messidor-2 is used to compare the performance of the model in a purely binary task (*referable / non-referable*).

Referable	Test Samples
No	1368
Yes	380

Inception-V3 was trained for the binary *referable / non-referable* DR diagnosis. Taking the output of the last average pooling layer of the Inception-V3 we got a 2048-dimensional vector representing each eye fundus image.

Quantum Measurement Regression Again, three hyper-parameters need to be chosen for the QMR stage: γ and D for the RFF layer and the number of eigen-components for the Quantum Measurement layer. We performed a random search for the QMR hyper-parameters fixing the Inception-V3 stage and generating 25 different random configurations. As result, we choose an embedding with 128 RFF components, 8 eigen-components and γ was set to 2^{-11}. For the loss function (Eq. 2), α was set at 0.6.

Baseline Similar to the baseline used in [40], an extension of the feature extractor model with two dense layers was set up as baseline for this work. The model, called *DLC-DR* hereafter, has a single neuron output and was trained end-to-end for classification. The Gaussian process approach corresponds to the model presented in [40]. And also, as we did for the PCa diagnosis case, we report results from the deep Gaussian process [7] and the DMKDC model[12].

5 Experimental Results and Discussion

To measure the performance of an ordinal regression method implies to take into account the severity of the misclassified samples. Although there is quite a variety of metrics, *Mean Absolute Error* (MAE) is currently a widely used measure in ordinal regression, both for evaluation and for the loss function of the models [11]. Therefore we also measured MAE on the test sets. As we also

want to measure the impact on the respective binary classifications, in each application a direct binarization of the results is made and compared with the performance reported by models trained for this purpose in the state of the art.

5.1 Prostate Cancer

WSI scores were summarized by a MV and PV. The prediction methods at WSI-level were also applied to the baseline models. In the dense layer classifiers from the softmax output, as in [41]. In the DMKDC, the prediction methods were easily applied because the model outputs a probability distribution. For GP and DGP only MV was calculated, since we have no access to an explicit discrete posterior distribution. The results are reported in Table 4 and Table 5. In terms

Table 4. Patch-level results of the dense layers classifier models DCL, Gaussian processes GP, DGP, and density matrix-based models DMKDC, DQOR.

Method	Accuracy	Macro F1	MAE
DLC-PCa [19]	**0.593**	**0.359**	0.698
GP [31]	0.399	0.255	0.777
DGP [7]	0.265	0.169	1.013
DMKDC [12]	0.584	0.377	0.717
DQOR	0.515	0.317	**0.6807**

of accuracy at the patch level, the DLC-PCa model obtained the highest results. This makes sense as this model is trained to optimize a categorical cross-entropy loss function. The difference with the regression approach is perceptible in the MAE, for which DQOR reached the best performance.

The best accuracy at the WSI level was also reached with the DLC-PCa model and with the DMKDC with probability vote. Regarding the regression performance, the DQOR obtained the lowest MAE at the WSI level, showing that the model takes advantage of the probability distributions and the inherent ordinality of the GS grades.

By directly categorizing the results in the test set, where a GS of 6 or 7 is marked as low, and a GS of 8, 9 or 10 as high, we binarized the results and calculated the accuracy to make a direct comparison with previous work using the same dataset. The results are reported in Table 6. The results reported in [15], [32] are obtained by means of CNN's directly trained with binary labels. The model presented in [19] is a multimodal approach, which uses the additional information obtained from the free text reports of the WSI's to enrich a model that makes inferences from visual information alone. Such training also uses binary labels.

Overall, this shows that in addition to performing better on the gradation task, our approach offers a consistent benefit for the subsequent binary task.

Table 5. WSI-level results. For each model, two summarization procedures are applied, majority vote (MV) and probability vote (PV).

Method	Accuracy	Macro F1	MAE
DLC-PCa MV [19]	**0.608**	0.354	0.7173
GP MV [31]	0.391	0.233	0.739
DGP MV [7]	0.174	0.059	0.935
DMKDC MV [12]	0.608	0.354	0.717
DQOR MV	0.587	**0.361**	0.695
DLC-PCa PV [19]	**0.608**	0.354	0.717
DMKDC PV [12]	**0.608**	0.354	0.717
DQOR PV	0.587	0.356	**0.652**

Table 6. Results at WSI-level of *low* risk vs *high* risk.

Method	Accuracy
Google LeNet [15]	0.7352
Modified AlexNet[32]	0.769
M-LSA [19]	0.770
DQOR	**0.782**

5.2 Diabetic Retinopathy

The results for DR grading on the EyePACS test set are reported in Table 7. Again, while the methods evaluated in this work generate a continuous prediction interpretable at five levels, binarization of these results can be done in a straightforward manner by defining a threshold on the prediction value. Although the sensitivity and specificity values will depend on this threshold, the ROC-AUC is independent of it and we report it in order to directly compare it with the state of the art. For Messidor-2 we only report binary classification performance in Table 8.

Table 7. Comparison on EyePACS test partition results. Sensitivity, specificity and AUC for binary classification and MAE for grading.

Description	Sensitivity	Specificity	AUC	MAE
DLC-DR [40]	0.7867	0.9643	0.9471	1.1011
Voets 2019 [43]	0.906	0.847	0.951	–
GP [40]	**0.9323**	0.9173	0.9769	0.7750
DGP [7]	0.3703	0.6196	0.4947	1.3566
DMKDC [12]	0.6473	0.8787	0.9135	0.4051
DQOR	0.8660	**0.9809**	**0.9805**	**0.2871**

Table 8. Comparison on Messidor-2 results. Sensitivity, specificity and AUC for binary classification.

Description	Sensitivity	Specificity	AUC
DLC-DR [40]	0.6105	**0.9715**	0.8624
Voets 2019 [43]	0.818	0.712	0.853
GP [40]	0.7237	0.8625	0.8787
DGP [7]	0.4026	0.5782	0.4960
DMKDC [12]	0.3894	0.5307	0.4581
DQOR	**0.8289**	0.9356	**0.9329**

It is evident that for both the grading task and the binary diagnosis task, our proposed DQOR improves the performance of the previous models, justifying once again the importance of using disease stage information.

5.3 Uncertainty Quantification

Beyond the classification and regression performance of the methods, DQOR allows an uncertainty quantification based on the variance of the predicted distribution for each sample. For the prostate cancer diagnosis, we analyzed the statistical behavior of the predicted variance on the test set at the WSI-level, grouping the samples according to whether or not they were correctly classified in the binary task (see Figure 5). A similar procedure was performed for the DR diagnosis (see Figure 6). As expected, DQOR predicts low uncertainty levels on well classified samples when compared with the miss-classified samples. In the case of DR diagnosis, it is remarkable the fact that, for different datasets (EyePACS and Messidor-2) the range of the variance is directly comparable, and the general statistical behavior is quite similar. This attribute provides the method with an interpretable means for the specialist, who may decide whether to trust or not in the model prediction.

6 Conclusions

In this work we present a novel pipeline for medical image analysis that combined the representational power of deep learning with the Quantum Measurement Regression method, which uses density matrices and random features to build a non-parametric density estimator.

We tested our approach in two different tasks: the diagnosis of prostate cancer and diabetic retinopathy diagnosis. In both cases the diagnosis is based on a gradation by progressive levels, although it is normal that in the end a binary diagnosis is taken into account, mainly related and justified by the prognosis of the disease. The training of the models was performed using the five available grades, and we report both regression and binary classification results.

Comparing with similar regression and classification schemes, the results show consistently that the DQOR allows to obtain better results in terms of

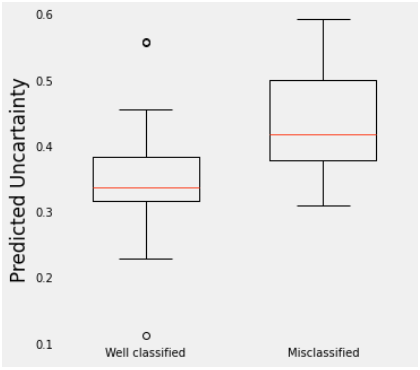

Fig. 5. Box plot of the predicted uncertainty on TCGA test samples at WSI-level, grouped by classification status on the low risk vs. high risk GS diagnosis task.

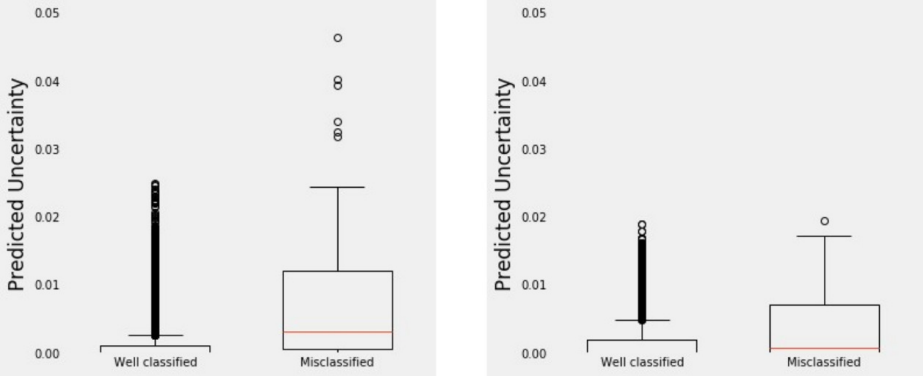

Fig. 6. Box plot of the predicted uncertainty on EyePACS test samples (left) and Messidor-2 (right), grouped by their classification status on the *referable / non-referable* diagnosis task.

MAE, which for medical applications implies a great advantage, given the sensitivity to the magnitude of classification errors that a purely categorical metric does not have. Furthermore, by directly binarizing the results, we show that a training using the information of the grades allows to improve the final binary classification performance.

Furthermore, unlike methods based solely on neural networks and other probabilistic models, DQOR predicts for each sample a discrete probability distribution over the range of labels. This allows for the robust integration of results in applications requiring patch-based analysis, and more importantly, it allows for uncertainty measurement to be involved during training. In test cases, we show that this uncertainty is significantly higher in misdiagnosed cases, and furthermore that the statistical behavior of this measurement is consistent across different datasets. This means that the method is able to tell us the level of confidence of its inference and can help in the identification of misclassified samples. This is a highly valued ability in medical applications, where the aim is to prevent false positives and especially false negatives in a diagnostic process.

Overall we demonstrate that unlike single deep learning architectures and standard classification models, the combination of deep CNNs and Quantum Measurement Regression allows us to use the ordinal information of a disease grades and provides a better theoretical framework to combine the patch-level inference into a single prediction and to quantify uncertainty in safety-critical applications.

References

1. Abràmoff, M.D., Folk, J.C., Han, D.P., Walker, J.D., Williams, D.F., Russell, S.R., Massin, P., Cochener, B., Gain, P., Tang, L., Lamard, M., Moga, D.C., Quellec, G., Niemeijer, M.: Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmology **131**(3), 351–357 (2013). https://doi.org/10.1001/jamaophthalmol.2013.1743
2. Adler, T.J., Ardizzone, L., Vemuri, A., Ayala, L., Gröhl, J., Kirchner, T., Wirkert, S., Kruse, J., Rother, C., Köthe, U., Maier-Hein, L.: Uncertainty-aware performance assessment of optical imaging modalities with invertible neural networks. International Journal of Computer Assisted Radiology and Surgery **14**(6), 997–1007 (2019). https://doi.org/10.1007/s11548-019-01939-9, https://doi.org/10.1007/s11548-019-01939-9
3. American Academy of Ophthalmology: International clinical diabetic retinopathy disease severity scale detailed table. International Council of Ophthalmology (2002)
4. Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, Â., Mendonça, A.M., Campilho, A.: DR—GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Medical Image Analysis **63** (2020). https://doi.org/10.1016/j.media.2020.101715
5. Beckham, C., Pal, C.: A simple squared-error reformulation for ordinal classification (Nips) (2016), http://arxiv.org/abs/1612.00775
6. Beckham, C., Pal, C.: Unimodal probability distributions for deep ordinal classification. 34th International Conference on Machine Learning, ICML 2017 **1**, 647–655 (2017)

7. Cutajar, K., Bonilla, E.V., Michiardi, P., Filippone, M.: Random feature expansions for Deep Gaussian Processes. 34th International Conference on Machine Learning, ICML 2017 **2**, 1467–1482 (2017)
8. Diabetic Retinopathy Detection of Kaggle: Eyepacs challenge. www.kaggle.com/c/diabetic-retinopathy-detection/data, accessed: 2019-10-15
9. Faraj, S.F., Bezerra, S.M., Yousefi, K., Fedor, H., Glavaris, S., Han, M., Partin, A.W., Humphreys, E., Tosoian, J., Johnson, M.H., Davicioni, E., Trock, B.J., Schaeffer, E.M., Ross, A.E., Netto, G.J.: Clinical validation of the 2005 isup gleason grading system in a cohort of intermediate and high risk men undergoing radical prostatectomy. PLoS ONE **11**(1), 1–13 (2016). https://doi.org/10.1371/journal.pone.0146189
10. Frank, E., Hall, M.: A simple approach to ordinal classification. In: De Raedt, L., Flach, P. (eds.) Machine Learning: ECML 2001. pp. 145–156. Springer Berlin Heidelberg, Berlin, Heidelberg (2001)
11. Garg, B., Manwani, N.: Robust deep ordinal regression under label noise. In: Pan, S.J., Sugiyama, M. (eds.) Proceedings of The 12th Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 129, pp. 782–796. PMLR, Bangkok, Thailand (18–20 Nov 2020), http://proceedings.mlr.press/v129/garg20a.html
12. González, F.A., Gallego, A., Toledo-Cortés, S., Vargas-Calderón, V.: Learning with Density Matrices and Random Features (2021), http://arxiv.org/abs/2102.04394
13. Gunawardhana, P.L., Jayathilake, R., Withanage, Y., Ganegoda, G.U.: Automatic Diagnosis of Diabetic Retinopathy using Machine Learning: A Review. Proceedings of ICITR 2020 - 5th International Conference on Information Technology Research: Towards the New Digital Enlightenment (2020). https://doi.org/10.1109/ICITR51448.2020.9310818
14. Gutiérrez, P.A., Pérez-Ortiz, M., Sánchez-Monedero, J., Fernández-Navarro, F., Hervás-Martínez, C.: Ordinal Regression Methods: Survey and Experimental Study. IEEE Transactions on Knowledge and Data Engineering **28**(1), 127–146 (2016). https://doi.org/10.1109/TKDE.2015.2457911
15. Jiménez del Toro, O., Atzori, M., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rönnquist, P., Müller, H.: Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score. Medical Imaging 2017: Digital Pathology **10140**, 101400O (2017). https://doi.org/10.1117/12.2255710
16. Karimi, D., Nir, G., Fazli, L., Black, P.C., Goldenberg, L., Salcudean, S.E.: Deep Learning-Based Gleason Grading of Prostate Cancer from Histopathology Images - Role of Multiscale Decision Aggregation and Data Augmentation. IEEE Journal of Biomedical and Health Informatics **24**(5), 1413–1426 (may 2020). https://doi.org/10.1109/JBHI.2019.2944643
17. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems **2017-December**(Nips), 5575–5585 (2017)
18. Khani, A.A., Fatemi Jahromi, S.A., Shahreza, H.O., Behroozi, H., Baghshah, M.S.: Towards Automatic Prostate Gleason Grading Via Deep Convolutional Neural Networks. 5th Iranian Conference on Signal Processing and Intelligent Systems, ICSPIS 2019 (December), 18–19 (2019). https://doi.org/10.1109/ICSPIS48872.2019.9066019
19. Lara, J.S., Contreras O., V.H., Otálora, S., Müller, H., González, F.A.: Multimodal latent semantic alignment for automated prostate tissue classification and

retrieval. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 572–581. Springer International Publishing, Cham (2020)

20. Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. Scientific Reports **7**(1), 1–14 (2017). https://doi.org/10.1038/s41598-017-17876-z
21. Li, H., Habes, M., Fan, Y.: Deep Ordinal Ranking for Multi-Category Diagnosis of Alzheimer's Disease using Hippocampal MRI data. arXiv (sep 2017), http://arxiv.org/abs/1709.01599
22. Li, Y., Huang, M., Zhang, Y., Chen, J., Xu, H., Wang, G., Feng, W.: Automated Gleason Grading and Gleason Pattern Region Segmentation Based on Deep Learning for Pathological Images of Prostate Cancer. IEEE Access **8**, 117714–117725 (2020). https://doi.org/10.1109/ACCESS.2020.3005180
23. Liu, J., Ren, Z.H., Qiang, H., Wu, J., Shen, M., Zhang, L., Lyu, J.: Trends in the incidence of diabetes mellitus: results from the Global Burden of Disease Study 2017 and implications for diabetes mellitus prevention. BMC Public Health **20**(1), 1–12 (2020). https://doi.org/10.1186/s12889-020-09502-x
24. Liu, X.: Ordinal Regression with Neuron Stick-breaking for Medical Diagnosis. Tech. rep. (2018)
25. Lucas, M., Jansen, I., Savci-Heijink, C.D., Meijer, S.L., de Boer, O.J., van Leeuwen, T.G., de Bruin, D.M., Marquering, H.A.: Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. Virchows Archiv **475**(1), 77–83 (2019). https://doi.org/10.1007/s00428-019-02577-x
26. Moccia, S., Wirkert, S.J., Kenngott, H., Vemuri, A.S., Apitz, M., Mayer, B., De Momi, E., Mattos, L.S., Maier-Hein, L.: Uncertainty-aware organ classification for surgical data science applications in laparoscopy. IEEE Transactions on Biomedical Engineering **65**(11), 2649–2659 (2018). https://doi.org/10.1109/TBME.2018.2813015
27. Nagpal, K., Foote, D., Liu, Y., Chen, P.H.C., Wulczyn, E., Tan, F., Olson, N., Smith, J.L., Mohtashamian, A., Wren, J.H., Corrado, G.S., MacDonald, R., Peng, L.H., Amin, M.B., Evans, A.J., Sangoi, A.R., Mermel, C.H., Hipp, J.D., Stumpe, M.C.: Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. npj Digital Medicine **2**(1), 1–10 (dec 2019). https://doi.org/10.1038/s41746-019-0112-2, https://doi.org/10.1038/s41746-019-0112-2
28. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output CNN for age estimation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2016-Decem, pp. 4920–4928. IEEE Computer Society (dec 2016). https://doi.org/10.1109/CVPR.2016.532
29. Perdomo, O., Gonzalez, F.: A Systematic Review of Deep Learning Methods Applied to Ocular Images. Ciencia e Ingenieria Neogranadina **30**(1) (2019)
30. Rahimi, A., Recht, B.: Random features for large-scale kernel machines. In: Advances in Neural Information Processing Systems 20 - Proceedings of the 2007 Conference (2009)
31. Rasmussen, C.E., Williams, C.K.I.: Gaussian processes for machine learning. The MIT Press (2006). https://doi.org/10.1142/S0129065704001899, http://www.gaussianprocess.org/gpml/chapters/RW.pdf

32. Ren, J., Hacihaliloglu, I., Singer, E.A., Foran, D.J., Qi, X.: Unsupervised Domain Adaptation for Classification of Histopathology Whole-Slide Images. Frontiers in Bioengineering and Biotechnology **7**(May), 1–12 (2019). https://doi.org/10.3389/fbioe.2019.00102
33. Sahran, S., Albashish, D., Abdullah, A., Shukor, N.A., Hayati Md Pauzi, S.: Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading. Artificial Intelligence in Medicine **87**, 78–90 (2018). https://doi.org/10.1016/j.artmed.2018.04.002, https://doi.org/10.1016/j.artmed.2018.04.002
34. Singh, A., Sengupta, S., Lakshminarayanan, V.: Explainable deep learning models in medical image analysis. Journal of Imaging **6**(6), 1–19 (2020). https://doi.org/10.3390/JIMAGING6060052
35. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Medical Image Analysis **67**, 101813 (2021). https://doi.org/10.1016/j.media.2020.101813, https://doi.org/10.1016/j.media.2020.101813
36. Stolte, S., Fang, R.: A survey on medical image analysis in diabetic retinopathy. Medical Image Analysis **64**, 101742 (2020). https://doi.org/10.1016/j.media.2020.101742, https://doi.org/10.1016/j.media.2020.101742
37. Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D.M., Bostwick, D.G., Evans, A.J., Grignon, D.J., Humphrey, P.A., Iczkowski, K.A., Kench, J.G., Kristiansen, G., van der Kwast, T.H., Leite, K.R., McKenney, J.K., Oxley, J., Pan, C.C., Samaratunga, H., Srigley, J.R., Takahashi, H., Tsuzuki, T., Varma, M., Zhou, M., Lindberg, J., Lindskog, C., Ruusuvuori, P., Wählby, C., Grönberg, H., Rantalainen, M., Egevad, L., Eklund, M.: Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. The Lancet Oncology **21**(2), 222–232 (2020). https://doi.org/10.1016/S1470-2045(19)30738-7
38. Sun, Y., Tang, J., Sun, Z., Tistarelli, M.: Facial Age and Expression Synthesis Using Ordinal Ranking Adversarial Networks. IEEE Transactions on Information Forensics and Security **15**, 2960–2972 (2020). https://doi.org/10.1109/TIFS.2020.2980792
39. Tian, L., Ma, L., Wen, Z., Xie, S., Xu, Y.: Learning Discriminative Representations for Fine-Grained Diabetic Retinopathy Grading (2020), http://arxiv.org/abs/2011.02120
40. Toledo-Cortés, S., De La Pava, M., Perdómo, O., González, F.A.: Hybrid Deep Learning Gaussian Process for Diabetic Retinopathy Diagnosis and Uncertainty Quantification. In: Ophthalmic Medical Image Analysis. OMIA 2020. Lecture Notes in Computer Science, vol 12069. Springer, Cham. pp. 206–215 (2020). https://doi.org/https://doi.org/10.1007/978-3-030-63419-321, https://doi.org/10.1007/978-3-030-63419-321
41. Tolkach, Y., Dohmgörgen, T., Toma, M., Kristiansen, G.: High-accuracy prostate cancer pathology using deep learning. Nature Machine Intelligence **2**(7), 411–418 (jul 2020). https://doi.org/10.1038/s42256-020-0200-7, https://www.nature.com/articles/s42256-020-0200-7
42. Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., Schön, T.: Evaluating model calibration in classification. In: Chaudhuri, K., Sugiyama, M. (eds.) Proceedings of Machine Learning Research. Proceedings of Ma-

chine Learning Research, vol. 89, pp. 3459–3467. PMLR (16–18 Apr 2019), http://proceedings.mlr.press/v89/vaicenavicius19a.html
43. Voets, M., Møllersen, K., Bongo, L.A.: Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. PLoS ONE **14**(6), 1–11 (2019). https://doi.org/10.1371/journal.pone.0217541
44. Voets, M., Møllersen, K., Bongo, L.A.: Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. PLoS ONE **14**(6), 1–11 (2019)
45. Wang, D., Foran, D.J., Ren, J., Zhong, H., Kim, I.Y., Qi, X.: Exploring automatic prostate histopathology image gleason grading via local structure modeling. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS **2015-Novem**, 2649–2652 (2015). https://doi.org/10.1109/EMBC.2015.7318936
46. Wells, J.A., Glassman, A.R., Ayala, A.R., Jampol, L.M., Bressler, N.M., Bressler, S.B., Brucker, A.J., Ferris, F.L., Hampton, G.R., Jhaveri, C., Melia, M., Beck, R.W.: Aflibercept, Bevacizumab, or Ranibizumab for Diabetic Macular Edema Two-Year Results from a Comparative Effectiveness Randomized Clinical Trial. Ophthalmology **123**(6), 1351–1359 (2016)