Computers in Biology and Medicine 145 (2022) 105472

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Grading diabetic retinopathy and prostate cancer diagnostic images with deep quantum ordinal regression

Santiago Toledo-Cortés [a], Diego H. Useche [a], Henning Müller [b], Fabio A. González [a,*]

[a] MindLab Research Group, Universidad Nacional de Colombia, Bogotá, Colombia
[b] Institute of Information Systems, HES-SO (University of Applied Sciences and Arts Western Switzerland), Sierre, Switzerland

ARTICLE INFO

Keywords:
Deep probabilistic learning
Density matrices
Diabetic retinopathy
Eye fundus images
Histopathology images
Ordinal regression
Prostate cancer
Quantum measurement
Uncertainty quantification

ABSTRACT

Although for many diseases there is a progressive diagnosis scale, automatic analysis of grade-based medical images is quite often addressed as a 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 labels and thus takes into account the order of grading scales used to assess the severity of different medical conditions. This paper presents a quantum-inspired deep probabilistic learning ordinal regression model for medical image diagnosis that takes advantage of the representational power of deep learning and the intrinsic ordinal information of disease stages. The method is evaluated on two different medical image analysis tasks: prostate cancer diagnosis and diabetic retinopathy grade estimation on eye fundus images. The experimental results show that the proposed method not only improves the diagnosis performance on the two tasks but also the interpretability of the results by quantifying the uncertainty of the predictions in comparison to conventional deep classification and regression architectures. The code and datasets are available at https://github.com/stoledoc/DQOR.

1. Introduction

The stages of a disease are not categorical. The degenerative process of a disease is not a discrete jump from one class to another but a progressive continuum [1]. These stages are therefore an attempt of the specialists to discretize a continuous behavior. While not completely accurate, this information is useful in the generation of automatic systems if a model with an appropriate descriptive capability is used. However, the information of the relative distance between the different grades of a disease is disregarded when a categorical classification model is used. The way to exploit this grading is therefore through a regression model. In addition, if a probabilistic regression model is implemented, the predictions can be interpreted as probability distributions over the range of the labels. Hence, one can infer the stage of the disease in a non-categorical way, providing more statistical information, for instance, the uncertainty of the predictions.

In the medical field, deep CNNs have been demonstrated to be effective at analyzing images and visual content of all kinds, from X-rays to diagnose osteoporosis [2], to MRIs to diagnose brain conditions [3]. Although many diseases present different stages on a progressive scale and in many cases this information is available, binary labels are usually favored [1]. However, two drawbacks arise from addressing the task of classifying grade-based medical images as a categorical problem with conventional neural networks: first, the ordinal information of the grades is not taken into account for the training process, and second, the predictions of the models, usually subject to a softmax activation function, cannot be interpreted as probability distributions [4].

In this paper, we present the Deep Quantum Ordinal Regressor (DQOR), a deep probabilistic model capable to combine a CNN with a differentiable probabilistic regression model, the Quantum Measurement Regression (QMR) [5,6]. DQOR is intended as a diagnostic support tool for the medical specialist which allows 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 image analysis, integrate patch posterior distributions into a single whole-slide image distribution using a simple, yet powerful probability-based strategy.
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.

* Corresponding author.
E-mail address: fagonzalezo@unal.edu.co (F.A. González).

https://doi.org/10.1016/j.compbiomed.2022.105472
Received 24 November 2021; Received in revised form 28 March 2022; Accepted 29 March 2022
Available online 9 April 2022

4. Improve the posterior prognosis-oriented binary diagnosis, based on an ordinal grade-label end-to-end training.

To show the effectiveness of our DQOR proposal, we test it on two grade-based diagnostic 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 [7]. 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 label the tissue regions with Gleason patterns on a scale from 1 to 5. The sum of the two most dominant Gleason patterns gives the final Gleason score. Hence, the Gleason score ranges from 2 to 10. However, in practice, the specialists only consider the highest five grades, from 6 to 10, since biopsies with a grade below 3 are not taken into account [8]. The higher the grade, the more advanced cancer. Although the automatic classification of PCa with CNNs has been widely studied, the usual approach has been as multi-class or binary classification of low risk (6–7 GS) vs high risk (8–10 GS) tasks [9,10].

Something similar happens with DR, which is a consequence of diabetes mellitus (DM), one of the most prevalent diseases worldwide [11]. Early DR diagnosis allows preventing most of the severe consequences of the disease, including complete blindness [12]. One method for early and effective diagnosis of DM consists of the inspection of the retinal tissue. This is made using an eye fundus image, an RGB photo of the inner back of the eyeball that allows detecting specific lesions that are direct consequences of the alterations caused by diabetes. These lesions include microaneurysms, neovascularization, hemorrhages, exudates, etc. By counting the number of these lesions, an ophthalmologist specializing in the retina can give a diagnosis of the disease, on a five-level severity scale from 0 to 4, being 0 a negative case of DR, and 4 a case of proliferative DR (see Fig. 1) [13]. Many approaches have treated the problem from a multi-class classification perspective, or as a binary task, where a diagnosis of 0 or 1 corresponds to a case of *non-referable* DR and a case of 2, 3, or 4 corresponds to a case of *referable* DR [14].

With DQOR we can predict the grades of the disorder, and also binarize the predictions. It is therefore possible to compare the performance of our model with models trained specifically for binary diagnosis.

This paper is organized as follows: Section 2 presents a brief overview of the related work, Section 3 presents the theoretical framework of the method, and Section 4 presents the experimental setup. To validate our approach, we compare the performance of our model with state-of-the-art deep learning-based models, and with various 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 this approach has occurred alongside the rest of machine learning, as an intermediate field between regression and classification. According to Gutierrez et al. [16], the taxonomy of previously proposed methods for ordinal regression are: first, *naïve* approaches, which are 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 [17], 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 might perform better than others, so there is not an optimal ordinal regression approach [16].

In the medical field, while it is true that there have been some applications of ordinal regression models, there is not a clear and well-defined trend. Recently, ordinal regression by binary classifiers has been applied to facial age estimation [18,19], and diagnosis of Alzheimer's disease [20], taking advantage of the inherent ordinal severity of brain degeneration.

In addition to the predictive value, for medical applications, it is desirable to obtain a probability distribution over the possible output stages. Furthermore, the distribution describing the probability of belonging to a disease stage should be unimodal, so it is expected to use unimodal distributions for ordinal classification tasks [21]. Various models have been proposed where predictions are forced to follow Poisson or binomial distributions over the possible outputs [21,22], showing that, when needed, the ordinal approach improves the results compared to the conventional cross-entropy approach.

Regarding our application of interest, most of the works for PCa have been focused on classifying Whole Slide Images (WSI) by *low* and *high* GS [9]. 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 Ref. [23], the authors classify 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. [24] uses a NASNetLarge CNN, and summarizes the GS of the patches by counting the probabilities per class. In Karimi et al. [25], they proposed training three CNNs for patches of different sizes and summarizing the probabilities by logistic regression.

Recently, however, there has been a growing interest in GS grading. Proof of this is the *Prostate cANcer graDe Assessment* (PANDA) Challenge [26], and the recently proposed CNN architectures, which include a combination of an atrous spatial pyramid pooling and a regular CNN [8],

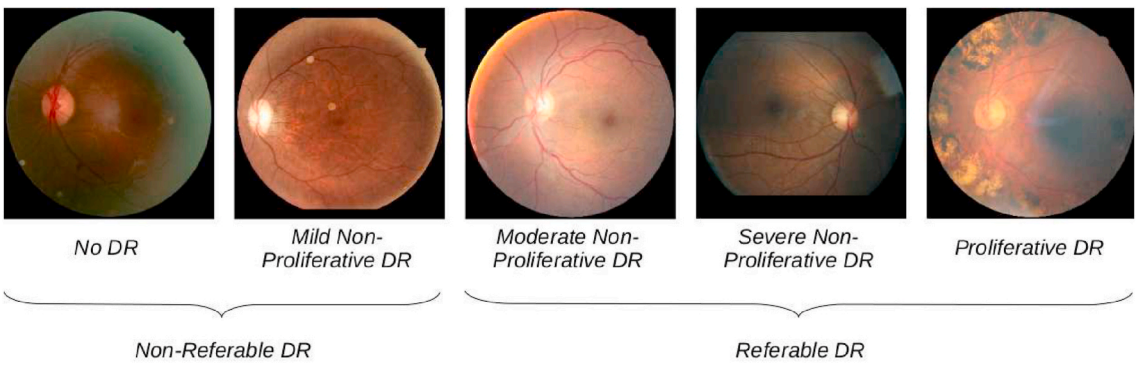

Fig. 1. Five possible grades for DR diagnosis. Healthy cases correspond to grade 0. Grades 0 and 1 correspond to *non-referable* DR cases, while grades 2, 3 and 4 correspond to *referable* DR. Samples extracted from EyePACS dataset [15].

an Inception-v3 CNN with a support vector machine (SVM) [27], and a DeepLabV3+ with a MobileNet as the backbone in Ref. [28]. In Ref. [29], the authors use an InceptionV3 with a k-nearest-neighbor classifier to summarize the patch-level predictions in a heatmap. In Ref. [30], the authors implemented in parallel a categorical and an ordinal classification for Gleason patterns, training similar models with different loss functions, from the same data features. However, they used a softmax to return probabilities which, as mentioned before, can not be interpreted directly as a probability distribution [4]. Other techniques for GS grading include, Support Vector Machine Feature-Recursive Feature Elimination [31], and learning features from *bag-of-words* features [32].

On the DR side, most works have been focused on a binary diagnosis based on deep neural networks [33,34]. Toledo-Cortés et al. [35] used the model proposed in Ref. [36] to extend a neural network-based binary classifier into a grading regressor through a Gaussian process. Tian et al. [37] also used a deep CNN as the backbone for a model trained to optimize a combination of a metric loss and a focal loss function for soft labels, aiming to use the ordinal information of the DR stages. Additionally, Teresa Araujo et al. [38] proposed the DR|GRADUATE, a deep learning-based model whose last layer had the same number of neurons as classes, and with a Gaussian filter applied to the output. The model was trained with a loss function that controlled both the entropy of the classification and the standard deviation of the distribution. This strategy allowed them to infer, in addition to the DR grade, the uncertainty of the prediction.

Uncertainty quantification in ordinal regression has been analyzed in many studies to obtain more interpretable models for cases where reliability is important to the end user [39]. Studies have been conducted on the uncertainty of machine learning algorithms for organ classification [40] and on the estimation of tissue parameters in the operating room [41]. Furthermore, estimating the uncertainty of a model's prediction has been of great interest because it reduces the consequences of the blind use of the model's inference [42]. This is particularly relevant in medical settings, where misdiagnoses can have serious consequences for patients. Correspondingly, Leibig et al. [43] analyzed 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. They claimed that Bayesian approaches perform better for uncertainty estimation and showed that uncertainty-aware decisions can improve the overall grading process.

Our method manages to capture many of the advantages of the previous methods, but with a stronger theoretical framework, and with a greater versatility of integration. For instance, DQOR is based on a probabilistic model, which allows the predictions to be actual probability distributions over the range of degrees, without the need to force them with the softmax or other activation functions. This in turn allows local predictions to be integrated into global predictions, in the case of PCa, and also allows the variance to be interpreted naturally as a measure of uncertainty. Finally, unlike classical probabilistic methods, DQOR can be trained with gradient descent, enabling its integration with conventional deep learning architectures.

3. Deep Quantum Ordinal Regressor

The overall architecture of the proposed Deep Quantum Ordinal Regressor (DQOR) is described in Fig. 2. We use a deep CNN as a feature extractor. The extracted features are then used as inputs for the QMR method [6]. QMR uses density matrices for regression problems and works as a density estimator. It requires an additional feature mapping from the inputs to get a quantum state-like representation. This is achieved using a random Fourier features approach [44]. The regressor yields a discrete posterior probability distribution from which we get the final grade prediction and a measure of the uncertainty.

3.1. Feature extraction

Medical and non-medical automatic image analysis relies on deep convolutional neural networks. The representational power of these models has shown remarkable results on computer vision and therefore we use them for feature extraction [45]. Regardless of the CNN architecture, these models conserve a basic structure: an input layer for the image, followed by a series of convolutional blocks and a pooling layer that summarizes all the information extracted from the convolutions. Usually, this layer is connected to a series of dense layers that perform the final classification of the image. Instead, we use the output of the pooling layer to feed the Quantum Measurement Regressor.

3.2. Random Fourier features

The random Fourier features (RFF) technique [44] 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. $w_1, \ldots, w_D \in \mathbb{R}^n$ from a probability distribution $p(w)$ given by the Fourier transform of $k(\mathbf{x} - \mathbf{y})$, and sampling i.i.d. $b_1, \ldots, b_D \in \mathbb{R}$ from a uniform distribution in $[0, 2\pi]$. In our context, the shift invariant kernel is the Radial Basis Function (RBF) given by, $k_{\mathrm{RBF}}(\mathbf{x} - \mathbf{y}) = e^{-\frac{1}{2\sigma^2}\left\|\mathbf{x}-\mathbf{y}\right\|^2}$, where σ 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 average pooling layer of the deep CNN module to a representation space that is suitable for the quantum measurement regression layer.

3.3. Quantum measurement regression (QMR)

QMR [6] is a differentiable probabilistic regression model that uses a

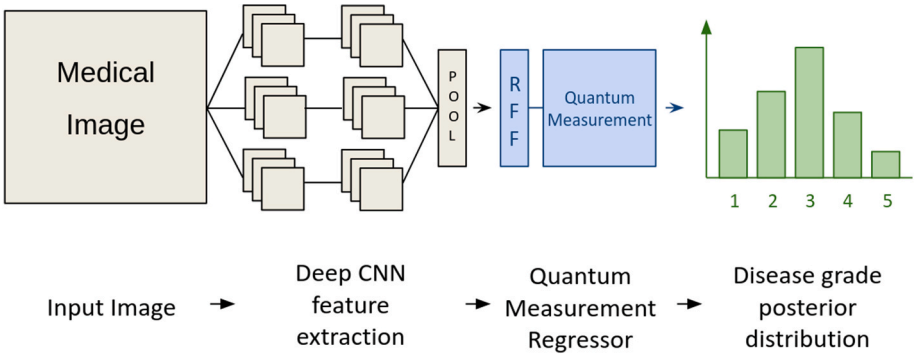

Fig. 2. Overview of the proposed DQOR method for medical image analysis. A deep CNN is used as feature extractor for the input image. Those features are the input for the QMR regressor model, which yields a posterior probability distribution over the possible grades of the disease.

trainable 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 $\pi = |\psi_x\rangle\langle\psi_x| \otimes \text{Id}_{\mathcal{H}_y}$ where $\text{Id}_{\mathcal{H}_y}$ is the identity operator in $\mathcal{H}_y$, the representation space of the labels. Inference is made by performing a quantum measurement on the training density matrix ρ_{train}:

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

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

A gradient-based optimization is allowed by a spectral decomposition of the density matrix,

$$\rho_{\text{train}} = V^{\dagger}\Lambda V, \tag{2}$$

in which the number of eigencomponents 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:

$$L = \sum (y - \hat{y})^2 + \alpha \sum_r \rho_{yrr}(\hat{y} - r)^2. \tag{3}$$

3.4. Patch-based analysis summarization

Patch-based image analysis is preferred or rather needed in some applications. This is the case of prostatic cancer diagnoses with WSI. We require the additional step of summarizing the predictions of the patches to reach a prediction of a whole slide image. The most straightforward procedure is the majority vote (MV), as reported in most previous works [9,46]. In the majority vote, the image's prediction is decided according to the grade with the highest number of predictions among the patches of the image. However, as in Ref. [24], DQOR admits 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,

$$P(r|I) = \frac{P(r,I)}{P(I)} = \frac{\sum_{i=1}^{n} P(r|p_i,I)P(p_i|I)P(I)}{P(I)} = \frac{1}{n}\sum_{i=1}^{n} P(r|p_i). \tag{4}$$

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

4. Experimental set up

The specific details of the experimental procedures are described below. The implementation was developed in Python, and the code is available at https://github.com/stoledoc/DQOR.

4.1. QMR hyperparameter optimization

As described in Section 3, the QMR layer of the DQOR requires five hyperparameters to be set before training. In addition to the usual learning rate, we have σ and D as hyperparameters of the RFF embedding, σ controls the spread of the objective Gaussian kernel that we try to approximate and D corresponds to the number of random Fourier features, which determines the dimension of the embedding space. Also, the training density matrix ρ_{train} (see Eq. (2)) depends on the number of eigencomponents n, and the loss function (see Eq. (5)) on the parameter α to control the variance of the predictions.

Due to the high number of possible hyperparameter combinations, we made a random search to look for the optimal setup. However, there are some important aspects to take into account to refine this process, which are presented below.

From a subset of the original dataset, we generated more than 3000 combinations of hyperparameters to train the model. In each case, we established an early stopping callback to halt the training process after 10 epochs with no improvement in the validation loss. We recorded the final *MAE* scores of the validation data set of all the resulting models. From this information, we made a statistical analysis of the hyperparameter sensibility of the model, by measuring the relative change of the *MAE* in comparison to the relative change of each hyperparameter. We looked at the density distribution of these combinations of hyperparameters as a function of the relative change of the *MAE* (see Fig. 3). We conclude that the more sensitive parameters of the QMR are learning rate and σ.

For the random search of the whole datasets, we set a range from 10^{-8} to 10^{-2} for the learning rate. Since σ measures the dispersion of the data, we took the mean of the pair-wise distances between the data samples and the mid-point of the data for the range of search. For the number of random features, D, we explored 2^7, 2^8, 2^9, 2^{10} and 2^{12}, taking into account that the output of the Xception and the Inception-V3 is a 2048-dimension vector. For the number of eigenvectors, we explored five different fractions of the chosen D: D, $D/2$, $D/4$, $D/8$, $D/16$. Finally, for the α parameter, we set the searched range from 0 to 1.

4.2. Prostate cancer

The setup of the DQOR applied to the PCa image analysis is described in Fig. 4.

4.2.1. Dataset

We used 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) [23]. To directly compare our results with the baseline [9], we used the same subset and partition for train and test, the details of the partition are presented in Table 1. The process to extract the patches from WSI is described in Ref. [23].

We used the images' patches to train the model. To obtain predictions at the level of WSIs, a process of summarization was carried out. Each patch was labeled with the same GS of the WSI from which it belongs. Although it is not clear that a GS can be assigned to a single patch, our methodology focused on showing the effectiveness of the regression approach by comparing it with previous works which use the labels of patches in the same manner, however, in theory, a GS can be set to each patch of a WSI.

4.2.2. Feature extraction

The model presented in Ref. [9] was used as a feature extractor. It is publicly available and consists of an Xception network trained on ImageNet 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 Ref. [9]. From the output of the last average pooling layer of the model, we got a 2048-dimensional vector representing each image patch.

4.2.3. Quantum measurement regression

For the QMR, hyper-parameter tuning of the model was performed by generating 25 different random configurations choosing the best combination. Henceforth, we created an embedding of 1024 RFF components with σ equals to 2^6. Also, the density matrix was trained with 32 eigenvalues. For the loss function (See eq. (3)) the value of 0.4 was selected for α, and the learning rate was set to 6×10^{-5}.

4.2.4. Baseline

We extended the feature extractor with a conventional feed-forward neural network as a baseline in this work. Called *DLC-PCa* hereafter, it

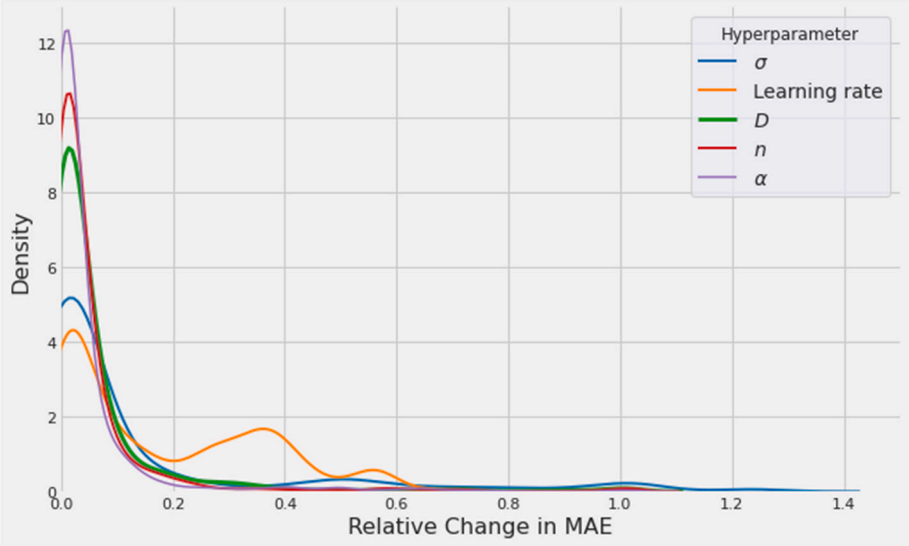

Fig. 3. Density plot of the ratio between the relative change of MAE and relative change of each hyperparameter. Although the mode of all distributions is close to zero, it can be noted that the variances of the learning rate and σ distributions are higher in comparison with the other three hyperparameters. This implies that the sensitivity of the model, measured against the variance of the MAE in a validation set, is higher in these two parameters.

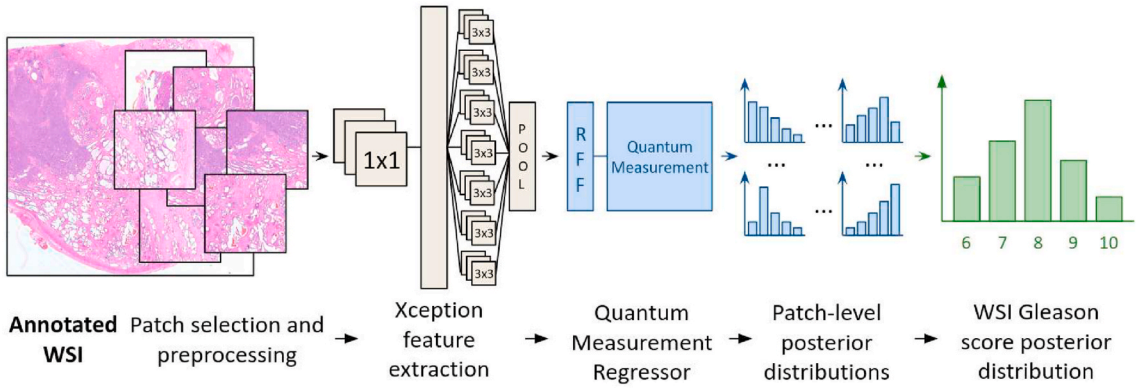

Fig. 4. Overview of the proposed DQOR method for prostate tissue grading. The Xception network was used as a feature extractor of the images' patches. Those features were the input for the QMR regressor model which yielded a posterior probability distribution by patch over the Gleason scores. Finally, those distributions were summarized into a single discrete probability distribution of the WSI.

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 Ref. [9].

Risk	Gleason Score	Train	Validation	Test
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

consists of 1024 neurons with ReLU as the activation function and a dropout of 0.2, followed by 5 neurons with a soft-max activation function for the output. The learning rate was set to 10^{-7}, as in the baseline [9]. We also explored two closely related methods to QMR: the Density Matrix Kernel Density Classification (DMKDC) [6] and the Gaussian process. DMKDC is a differentiable classification method, which applies the 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. A Gaussian process (GP) [47] is a powerful Bayesian approach to regression problems. Through 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 Gaussian kernel. The prediction process consists of marginalizing the learned Gaussian distribution from which the mean would be the actual predicted value and its standard deviation an indicator of the uncertainty. We also explored deep Gaussian processes (DGP) [48], which also use 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^{-7} in a single layer schema.

4.3. Diabetic retinopathy

The setup of the DQOR applied to eye fundus images' analysis is presented in Fig. 5.

4.3.1. Datasets

To directly compare the DQOR performance with the baselines of the state-of-the-art, we worked with EyePACS [15] and Messidor-2 [49] datasets. EyePACS is one of the largest publicly available datasets of eye fundus images. Each sample is labeled as one grade of the five grade

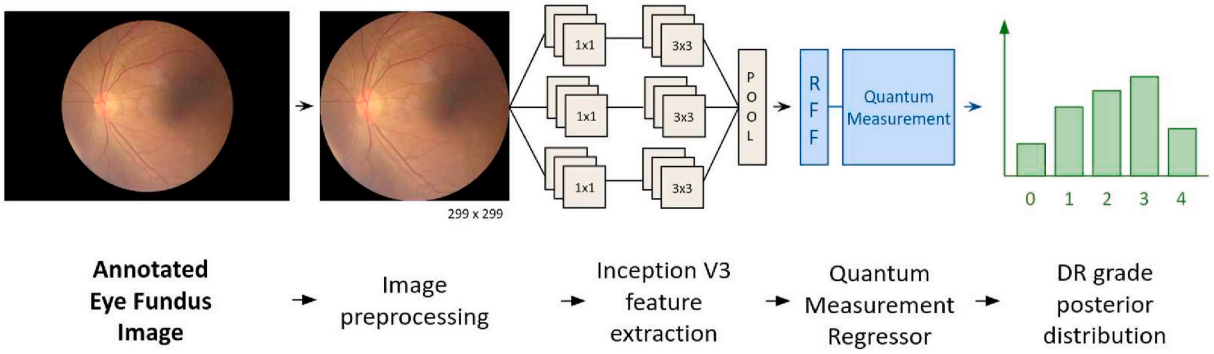

Fig. 5. Overview of the DQOR for diabetic retinopathy grading. An Inception-V3 network was used as feature extractor for the eye fundus image. These features were the input for the QMR regressor model, which yielded a posterior probability distribution over the DR grades.

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. To compare our method with our baselines [35,50], we kept the same partition for training and testing, which are described in Table 2. This data configuration of EyePACS has, however, a drawback: in the test data set there are no samples of DR grades 2 and 3. This is justifiable in the binary context for which it was originally designed, but it is not a fair evaluation for the ordinal classification case. Therefore, we set up a second partition for EyePACS, called EyePACS-b, based on the former one, but moving some samples from the train set to the test set. The details of the EyePACS-b partition are described in Table 3.

Regarding Messidor 2, it is a standard dataset in the field for testing. It consists of 1748 eye fundus images. While DR grades are not provided, we used them to show the effectiveness of our proposal for a *referable/non-referable* diagnosis. Details of Messidor-2 are described in Table 4.

4.3.2. Feature extraction

We used the model presented in Ref. [35] as a feature extractor, which is also publicly available and consists of an Inception-V3 network trained on ImageNet and fine-tuned on eye fundus images with the EyePACS train partition. This network had already been used as a feature extractor in a model for 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 Inception-V3 was trained for the binary *referable/non-referable* DR diagnosis. The same training setup was used to train an Inception-V3 on EyePACS-b. From the output of the last average pooling layer of the Inception-V3, we got a 2048-dimensional vector representing each eye fundus image.

4.3.3. Quantum measurement regression

We performed a random search for the QMR hyper-parameters fixing the Inception-V3 stage and generating 25 different random configurations. As result, we chose an embedding of 128 RFF components, 8 eigencomponents, and σ was set to 2^5. For the loss function (Eq. (3)), α was set at 0.6, optimized with a learning rate of 7×10^{-5}.

Table 2
Details of the subset and final partition of the EyePACS dataset used for training and testing. This partition is the same used in Ref. [35] and in Ref. [50].

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

Table 3
Details of the subset and final partition of the EyePACS-b partition used for a fair ordinal regression evaluation with samples in all the grades, in contrast to the partition used in Ref. [35] and in Ref. [50].

Referable	Grade	Train	Test
No	0	37209	7407
No	1	3479	689
Yes	2	10298	2575
Yes	3	1637	409
Yes	4	1220	694

Table 4
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	Total
No	1368
Yes	380

4.3.4. Baseline

Similar to the baseline used for the PCa grading case, an extension of the feature extractor model with two dense layers was set up as a baseline for this task (called *DLC-DR* hereafter). Correspondingly, we report the results from the deep Gaussian process [48] and the DMKDC model [6]. We also present the results of the Gaussian process approach proposed in Ref. [35].

5. Experimental results and discussion

To measure the performance of an ordinal regression method requires taking into account the severity of misclassified samples. Usual categorical classification metrics, such as accuracy or F1-score, are not appropriate to estimate the actual performance of an ordinal classification. For example, given a sample whose actual label is grade 3, it is more severe if a model classifies the sample as grade 1 than grade 2. The separation between the predictions and the actual labels of the models is especially relevant in the medical field. Therefore, it is required a metric that quantifies the magnitude of the misclassification error. Between all the possible metrics, *Mean Absolute Error* (MAE) is currently one widely used measure in ordinal regression, both for evaluation and the loss function of the models [51]. The MAE can be computed by,

$$\frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|, \tag{5}$$

where m is the number of test samples, y_i is the true label, and $\hat{y}_i$ is the predicted value. MAE is a metric that penalizes misclassifications

according to their distance to the true labels. Therefore, in addition to the categorical classification metrics, we also measured and reported MAE on the test data sets.

Finally, since it is desirable to measure the performance of the models for the binary classification task, we binarized the regressor predictions and compared this strategy with state-of-the-art models which were built for this specific purpose. Accuracy was used to measure the performance in the PCa case, and sensitivity, specificity, and AUC for the DR case.

5.1. Prostate cancer

WSI scores were summarized utilizing MV and PV. The prediction methods at the WSI level were also applied to the baseline models. In the dense layer classifiers, the summarization was made from the softmax output, as in Ref. [24]. In the DMKDC, the summarization 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 at patch-level and at WSI-level are reported in Table 5 and Table 6 respectively.

In terms of multi-class accuracy at the patch level, the DLC-PCa model obtained the highest results. This was expected since this model is trained to optimize the categorical cross-entropy loss function. The difference with the regression approach is noticeable in the MAE, for which DQOR reached the best performance. At the WSI level, the best multi-class accuracy 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.

In general, it should be highlighted that higher performances in the categorical classification do not imply higher performances for ordinal classification. The difference between the model with the highest accuracy and the model with the lowest MAE is shown in Fig. 6. The DQOR confusion matrix indeed presents a higher concentration of samples around the diagonal, showing that the model takes advantage of the probability distributions and the inherent ordering of the GS grades.

By considering the predicted GS of 6 or 7 as *low* GS, and the predicted GS of 8, 9, or 10 as *high* GS, we binarized the results and computed the accuracy to make a direct comparison with previous works using the same dataset. The results are reported in Table 7.

The results reported in Refs. [23,52] were obtained by training the binary labels of the WSIs with CNNs. The model presented in Ref. [9] is a multimodal approach, which used text reports as additional information to enrich the predictions of the WSIs, this model makes inferences from visual information alone, and also used binary labels. We can see that DQOR reached the highest binary accuracy and hence, it performed better on the gradation task. This approach was beneficial for the posterior binarization of the model.

5.2. Diabetic retinopathy

The results for DR grading on the EyePACS test set are reported in Table 8. As previously mentioned, the methods tested in this work generated a prediction for a five-grade range, and a binarization of the results was performed by defining a threshold on the predicted value. Henceforth, we report the ROC-AUC score to directly compare it with the state of the art (see Fig. 7 and Fig. 8). For Messidor-2 we only report the binary classification performance in Table 9.

Table 5
Patch-level multiclass results of the dense layers classifier model DCL-PCa, Gaussian process GP, DGP, and density matrix-based models DMKDC, DQOR.

Method	Accuracy	Macro F1	MAE
DLC-PCa [9]	**0.593**	**0.359**	0.698
GP [47]	0.399	0.255	0.777
DGP [48]	0.265	0.169	1.013
DMKDC [6]	0.584	0.377	0.717
DQOR	0.515	0.317	**0.6807**

Table 6
WSI-level results. For each model, two summarization procedures were applied, majority vote (MV) and probability vote (PV).

Method	Accuracy	Macro F1	MAE
DLC-PCa MV [9]	**0.608**	0.354	0.7173
GP MV [47]	0.391	0.233	0.739
DGP MV [48]	0.174	0.059	0.935
DMKDC MV [6]	0.608	0.354	0.717
DQOR MV	0.587	**0.361**	0.695
DLC-PCa PV [9]	**0.608**	0.354	0.717
DMKDC PV [6]	**0.608**	0.354	0.717
DQOR PV	0.587	0.356	**0.652**

For the ordinal regression task, DQOR was the best performing model according to MAE. Furthermore, for both partitioned datasets, it reported the highest AUC for the binary *referable/non-referable* task.

Also, the results of DQOR using the EyePACS-b partition are reported in Table 10 and the confusion matrix is shown in Fig. 9, next to the confusion matrix for DLC-DR. DQOR had the lowest MAE and the highest AUC. Note that DLC-DR was the model with the highest sensitivity, and the second-highest AUC, just below DQOR. However, it was not the model with the second-lowest MAE. The confusion matrices show again that the results of the DQOR had a lower dispersion around the diagonal, which improved the results for the subsequent binary task.

In general, it is noticeable that for both the ordinal and the binary classification tasks, our proposed DQOR improves the performance of the previous models, which justifies once again the importance of using the different stages of the disease for an automatic diagnosis.

5.3. Uncertainty quantification

In addition to the performances of the method on the classification and regression tasks, DQOR offers an uncertainty quantification based on the variance of the predicted distribution for each sample. For the PCa diagnosis, we analyzed the statistical behavior of the predicted variance on the test data set at the WSI level, grouping the samples according to whether or not they were correctly classified on the binary task. Fig. 10 shows boxplots of the predicted variance for each group. A similar procedure was performed for the DR diagnosis (see Fig. 11). As expected, DQOR predicts low uncertainties on well-classified samples in comparison with miss-classified samples. For the case of DR diagnosis, it is remarkable that for the EyePACS and Messidor-2 datasets the range of the variances are directly comparable, and they have similar statistical behavior. The uncertainty quantification provides the specialist with a more interpretable result, from which he may decide whether to trust or not on the model's prediction.

6. Conclusions

In this work, we presented a novel method for grade-based medical image analysis. Intended as a diagnostic support tool for the practitioner, the method combines the representational power of deep learning with the Quantum Measurement Regression method [6], which uses density matrices and random features to build a density estimator.

We tested our approach in two different tasks: the diagnosis of prostate cancer and diabetic retinopathy. In both cases, the diagnosis was based on a gradation by progressive levels. The training of the models was performed using the five available grades, and we reported the results for both ordinal regression and binary classification tasks. The latter were obtained by direct re-categorization over the regressor's predictions.

Compared with similar regression and classification methods, the results show that the while DQOR does not guarantee a better multiclass accuracy, it consistently allows obtaining better results in terms

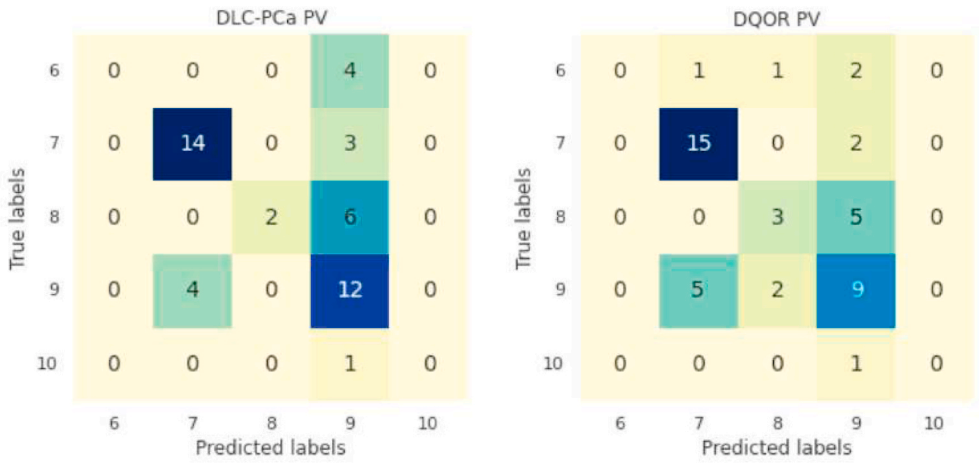

Fig. 6. Confusion matrices of the WSI grade predictions for DLC-PCa (left) and for DQOR (right) in the TCGA test partition. WSI prediction is obtained using the probability vote.

Table 7
Results at WSI-level of *low* risk vs *high* risk.

Method	Accuracy
Google LeNet [23]	0.7352
Modified AlexNet [52]	0.769
M-LSA [9]	0.770
DQOR	**0.782**

Table 8
Comparison on EyePACS test partition results. Sensitivity, specificity and AUC for binary classification and MAE for grading.

Description	Sensitivity	Specificity	AUC	MAE
DLC-DR [35]	0.7867	0.9643	0.9471	0.3166
Voets et al. (2019) [36]	0.906	0.847	0.951	–
GP [35]	**0.9323**	0.9173	0.9769	0.7750
DGP [48]	0.3703	0.6196	0.4947	1.3566
DMKDC [6]	0.6473	0.8787	0.9135	0.4051
DQOR	0.8660	**0.9809**	**0.9805**	**0.2871**

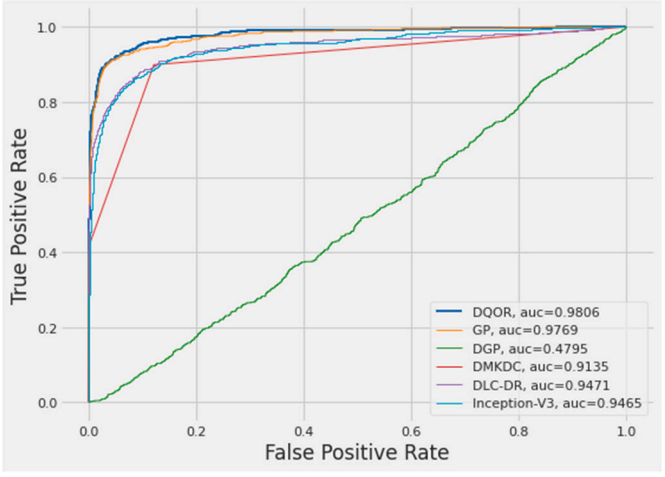

Fig. 7. ROC curves plot for EyePACS test set.

of MAE, which can be advantageous in medical applications, given the sensitivity to the magnitude of classification errors that a purely categorical metric does not have. Furthermore, by directly binarizing the results, we showed that training the models with the information of the

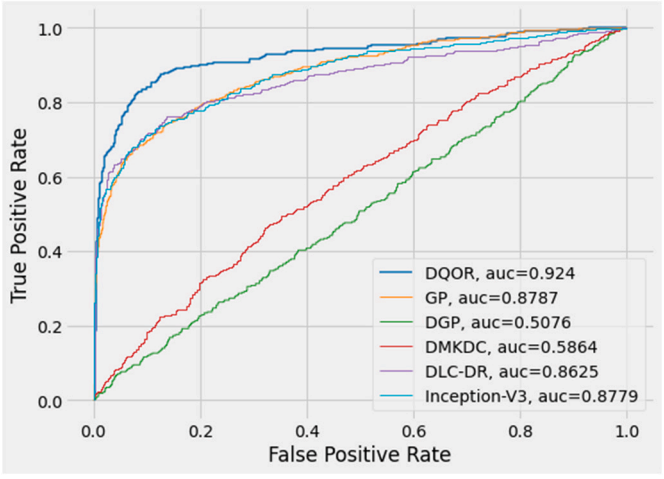

Fig. 8. ROC curves plot for Messidor 2.

Table 9
Comparison on Messidor-2 results. Sensitivity, specificity and AUC for binary classification.

Description	Sensitivity	Specificity	AUC
DLC-DR [35]	0.6105	**0.9715**	0.8624
Voets 2019 [36]	0.818	0.712	0.853
GP [35]	0.7237	0.8625	0.8787
DGP [48]	0.4026	0.5782	0.4960
DMKDC [6]	0.5906	0.5316	0.5864
DQOR	0.7974	0.9291	**0.9239**

Table 10
Comparison on EyePACS-b test partition results. Sensitivity, specificity and AUC for binary classification and MAE for grading.

Description	Sensitivity	Specificity	AUC	MAE
DLC-DR [35]	**0.9517**	0.7308	0.9363	0.4702
GP [35]	0.9454	0.5903	0.8385	0.5939
DGP [48]	0.2504	0.7510	0.5018	1.4954
DMKDC [6]	0.9166	0.8121	0.8736	0.4164
DQOR	0.9152	**0.8461**	**0.9438**	**0.3872**

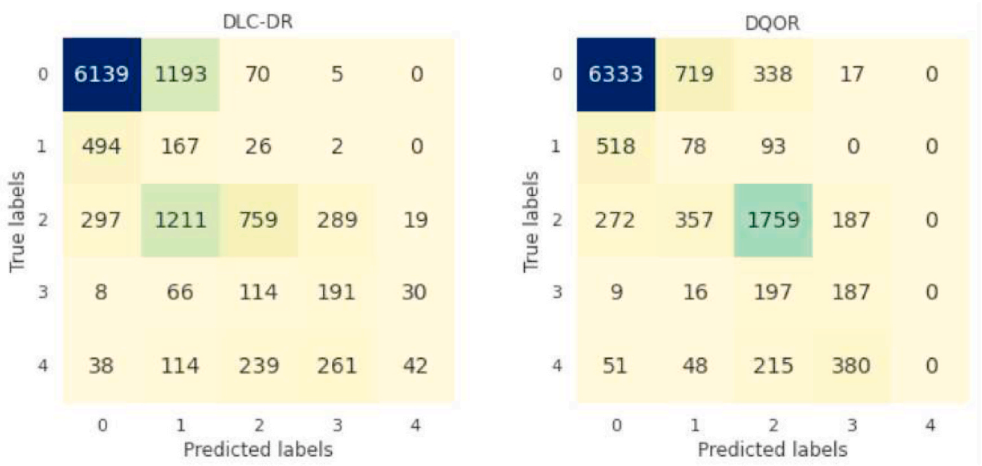

Fig. 9. Confusion matrices of the predictions of DLC-DR (left) and DQOR (right) in the EyePACS-b test partition.

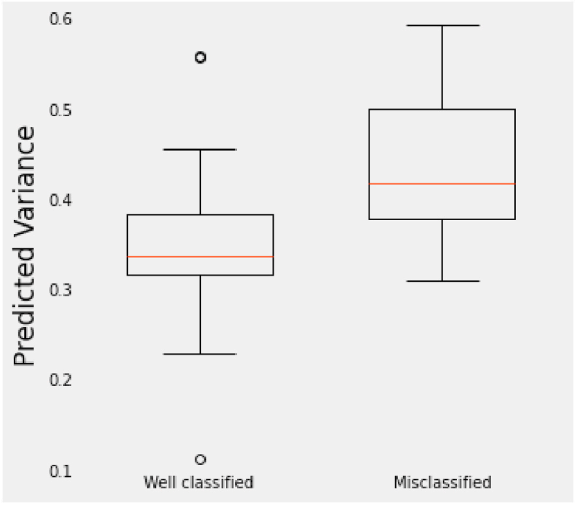

Fig. 10. Box plot of the predicted variance on TCGA test samples at WSI-level, grouped by classification status on the low risk vs. high risk GS diagnosis task.

grades improves the performance of the binary classification.

It should be highlighted that in these two medical applications the labels are presented on a progressive scale. Namely, the method takes advantage of the ordinal relationship of the labels, which is absent in the purely categorical tasks. In conventional multiclass classification problems the method is not expected to show improvements, on the contrary imposing an artificial order on the labels may negatively impact the 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 enables a robust integration of the results of patch-level images to a prediction on a whole slide image and offers the uncertainty of the prediction. In test cases, we showed that this uncertainty is significantly lower on well-classified samples in comparison to misdiagnosed samples and that the statistical behavior of this measurement is consistent across different datasets. This implies that the method is able to provide the level of confidence of its inference which can support the identification of misclassified samples. While this may require further research and statistical analysis, it is a highly valued feature for medical applications, where the goal is to prevent false positives and especially false negatives in a diagnostic process.

Overall, we demonstrated that unlike deep learning architectures and standard classification models, the combination of deep CNNs and Quantum Measurement Regression allows us to take advantage of the ordinal information of the stages of a disease in a probabilistic manner. This provides a better theoretical framework to deal with patch-based analysis, improves the performance in the binary prognosis-oriented diagnosis, and provides tools to quantify the uncertainty of the model for safety-critical applications.

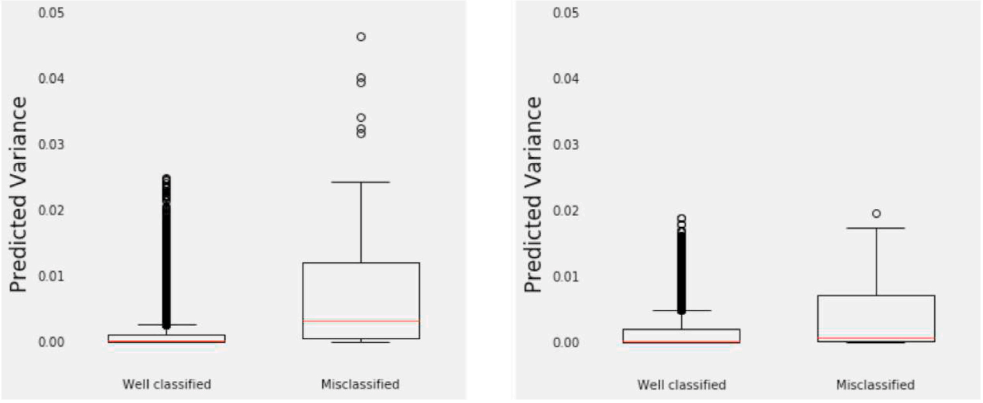

Fig. 11. Box plot of the predicted variance on EyePACS test samples (left) and Messidor-2 (right), grouped by their classification status on the *referable/non-referable* diagnosis task.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by a Google Research Award and by the Colciencias project number 1101-807-63563.

References

[1] X. Liu, Ordinal regression with neuron stick-breaking for medical diagnosis, Tech. Rep. (2018).
[2] K.S. Lee, S.K. Jung, J.J. Ryu, S.W. Shin, J. Choi, Evaluation of transfer learning with deep convolutional neural networks for screening osteoporosis in dental panoramic radiographs, J. Clin. Med. 9 (2) (2020), https://doi.org/10.3390/jcm9020392.
[3] D. Zhang, Y. Wang, L. Zhou, H. Yuan, Multimodal classification of Alzheimer's disease and mild cognitive impairment, Neuroimage (2011), https://doi.org/10.1038/jid.2014.371 arXiv:NIHMS150003.
[4] J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, T. Schön, Evaluating model calibration in classification, in: K. Chaudhuri, M. Sugiyama (Eds.), Proceedings of Machine Learning Research, vol. 89, PMLR, 2019, pp. 3459–3467. URL, http://proceedings.mlr.press/v89/vaicenavicius19a.html.
[5] F.A. González, V. Vargas-Calderón, H. Vinck-Posada, Classification with quantum measurements, J. Phys. Soc. Jpn. 90 (4) (2021) 44002, https://doi.org/10.7566/JPSJ.90.044002, arXiv.
[6] F. A. González, A. Gallego, S. Toledo-Cortés, V. Vargas-Calderón, Learning with Density Matrices and Random FeaturesarXiv:2102.4394. URL http://arxiv.org/abs/2102.04394.
[7] S.F. Faraj, S.M. Bezerra, K. Yousefi, H. Fedor, S. Glavaris, M. Han, A.W. Partin, E. Humphreys, J. Tosoian, M.H. Johnson, E. Davicioni, B.J. Trock, E.M. Schaeffer, A.E. Ross, G.J. Netto, 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) (2016) 1–13, https://doi.org/10.1371/journal.pone.0146189.
[8] Y. Li, M. Huang, Y. Zhang, J. Chen, H. Xu, G. Wang, W. Feng, Automated gleason grading and gleason pattern region segmentation based on deep learning for pathological images of prostate cancer, IEEE Access 8 (2020) 117714–117725, https://doi.org/10.1109/ACCESS.2020.3005180.
[9] J.S. Lara, V.H. Contreras O, S. Otálora, H. Müller, F.A. González, Multimodal latent semantic alignment for automated prostate tissue classification and retrieval, in: A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M.A. Zuluaga, S.K. Zhou, D. Racoceanu, L. Joskowicz (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing, Cham, 2020, pp. 572–581.
[10] P. Ström, K. Kartasalo, H. Olsson, L. Solorzano, B. Delahunt, D.M. Berney, D. G. Bostwick, A.J. Evans, D.J. Grignon, P.A. Humphrey, K.A. Iczkowski, J.G. Kench, G. Kristiansen, T.H. van der Kwast, K.R. Leite, J.K. McKenney, J. Oxley, C.C. Pan, H. Samaratunga, J.R. Srigley, H. Takahashi, T. Tsuzuki, M. Varma, M. Zhou, J. Lindberg, C. Lindskog, P. Ruusuvuori, C. Wählby, H. Grönberg, M. Rantalainen, L. Egevad, M. Eklund, Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study, Lancet Oncol. 21 (2) (2020) 222–232, https://doi.org/10.1016/S1470-2045(19)30738-7.
[11] J. Liu, Z.H. Ren, H. Qiang, J. Wu, M. Shen, L. Zhang, J. Lyu, Trends in the incidence of diabetes mellitus: results from the Global Burden of Disease Study 2017 and implications for diabetes mellitus prevention, BMC Publ. Health 20 (1) (2020) 1–12, https://doi.org/10.1186/s12889-020-09502-x.
[12] J.A. Wells, A.R. Glassman, A.R. Ayala, L.M. Jampol, N.M. Bressler, S.B. Bressler, A. J. Brucker, F.L. Ferris, G.R. Hampton, C. Jhaveri, M. Melia, R.W. Beck, Aflibercept, Bevacizumab, or Ranibizumab for diabetic Macular Edema two-year results from a comparative effectiveness randomized clinical trial, Ophthalmology 123 (6) (2016) 1351–1359.
[13] American Academy of Ophthalmology, International Clinical Diabetic Retinopathy Disease Severity Scale Detailed Table, International Council of Ophthalmology.
[14] S. Stolte, R. Fang, A survey on medical image analysis in diabetic retinopathy, Med. Image Anal. 64 (2020), 101742, https://doi.org/10.1016/j.media.2020.101742.
[15] Diabetic retinopathy detection of Kaggle, Eyepacs challenge. www.kaggle.com/c/diabetic-retinopathy-detection/data. (Accessed 15 October 2019).
[16] P.A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, C. Hervás-Martínez, Ordinal regression methods: survey and experimental study, IEEE Trans. Knowl. Data Eng. 28 (1) (2016) 127–146, https://doi.org/10.1109/TKDE.2015.2457911.
[17] E. Frank, M. Hall, A simple approach to ordinal classification, in: L. De Raedt, P. Flach (Eds.), Machine Learning: ECML 2001, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 145–156.
[18] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output CNN for age estimation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem, IEEE Computer Society, 2016, pp. 4920–4928, https://doi.org/10.1109/CVPR.2016.532.
[19] Y. Sun, J. Tang, Z. Sun, M. Tistarelli, Facial age and expression synthesis using ordinal ranking adversarial networks, IEEE Trans. Inf. Forensics Secur. 15 (2020) 2960–2972, https://doi.org/10.1109/TIFS.2020.2980792.
[20] H. Li, M. Habes, Y. Fan, Deep ordinal ranking for multi-category diagnosis of Alzheimer's disease using hippocampal MRI data, arXivarXiv:1709.1599. URL http://arxiv.org/abs/1709.01599.
[21] C. Beckham, C. Pal, Unimodal probability distributions for deep ordinal classification, 34th International Conference on Machine Learning, arXiv:1705.05278, ICML 1 (2017) 647–655, 2017.
[22] C. Beckham, C. Pal, A simple squared-error reformulation for ordinal classification (Nips), arXiv:1612.00775. URL, http://arxiv.org/abs/1612.00775.
[23] O. Jiménez del Toro, M. Atzori, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, P. Rönnquist, H. Müller, Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score, in: Medical Imaging 2017: Digital Pathology 10140, 2017, 101400O, https://doi.org/10.1117/12.2255710.
[24] Y. Tolkach, T. Dohmgörgen, M. Toma, G. Kristiansen, High-accuracy prostate cancer pathology using deep learning, Nat. Mach. Intell. 2 (7) (2020) 411–418, https://doi.org/10.1038/s42256-020-0200-7. URL, https://www.nature.com/articles/s42256-020-0200-7.
[25] D. Karimi, G. Nir, L. Fazli, P.C. Black, L. Goldenberg, S.E. Salcudean, Deep learning-based gleason grading of prostate cancer from histopathology images - role of multiscale decision aggregation and data augmentation, IEEE J. Biomed. Health Inform. 24 (5) (2020) 1413–1426, https://doi.org/10.1109/JBHI.2019.2944643.
[26] W. Bulten, G. Litjens, H. Pinckaers, P. Ström, M. Eklund, K. Kartasalo, M. Demkin, S. Dane, The PANDA Challenge: Prostate cANcer graDe Assessment Using the Gleason Grading System, Mar, 2020, https://doi.org/10.5281/zenodo.3715938.
[27] M. Lucas, I. Jansen, C.D. Savci-Heijink, S.L. Meijer, O.J. de Boer, T.G. van Leeuwen, D.M. de Bruin, H.A. Marquering, Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies, Virchows Arch. 475 (1) (2019) 77–83, https://doi.org/10.1007/s00428-019-02577-x.
[28] A.A. Khani, S.A. Fatemi Jahromi, H.O. Shahreza, H. Behroozi, M.S. Baghshah, 2019, in: Towards Automatic Prostate Gleason Grading via Deep Convolutional Neural Networks, 5th Iranian Conference on Signal Processing and Intelligent Systems, ICSPIS, 2019, pp. 18–19, https://doi.org/10.1109/ICSPIS48872.2019.9066019. December.
[29] K. Nagpal, D. Foote, Y. Liu, P.H.C. Chen, E. Wulczyn, F. Tan, N. Olson, J.L. Smith, A. Mohtashamian, J.H. Wren, G.S. Corrado, R. MacDonald, L.H. Peng, M.B. Amin, A.J. Evans, A.R. Sangoi, C.H. Mermel, J.D. Hipp, M.C. Stumpe, Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer, npj Digit. Med. 2 (1) (2019) 1–10, https://doi.org/10.1038/s41746-019-0112-2, arXiv:1811.06497.
[30] H. Huang, H. Situ, S. Zheng, Bidirectional information flow quantum state tomography, Chin. Phys. Lett. 38 (4) (2021) 1–6, https://doi.org/10.1088/0256-307X/38/4/040303, arXiv:2103.16781.
[31] S. Sahran, D. Albashish, A. Abdullah, N.A. Shukor, S. Hayati Md Pauzi, Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading, Artif. Intell. Med. 87 (2018) 78–90, https://doi.org/10.1016/j.artmed.2018.04.002.
[32] D. Wang, D.J. Foran, J. Ren, H. Zhong, I.Y. Kim, X. Qi, Exploring automatic prostate histopathology image gleason grading via local structure modeling, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2015-Novem, 2015, pp. 2649–2652, https://doi.org/10.1109/EMBC.2015.7318936.
[33] O. Perdomo, F. Gonzalez, A systematic review of deep learning methods applied to ocular images, Ciencia e Ingenieria Neogranadina 30 (1) (2020).
[34] J.Y. Choi, T.K. Yoo, J.G. Seo, J. Kwak, T.T. Um, T.H. Rim, Multi-categorical deep learning neural network to classify retinal images: a pilot study employing small database, PLoS One 12 (11) (2017) 1–16, https://doi.org/10.1371/journal.pone.0187336.
[35] S. Toledo-Cortés, M. De La Pava, O. Perdómo, F.A. González, Hybrid deep learning Gaussian process for diabetic retinopathy diagnosis and uncertainty quantification, arXiv:2007.14994, in: Ophthalmic Medical Image Analysis. OMIA 2020. Lecture Notes in Computer Science, vol. 12069Springer, Cham, 2020, pp. 206–215, https://doi.org/10.1007/978-3-030-63419-321. URL.
[36] M. Voets, K. Møllersen, L.A. Bongo, 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) (2019) 1–11, https://doi.org/10.1371/journal.pone.0217541, arXiv:1803.4337.
[37] L. Tian, L. Ma, Z. Wen, S. Xie, Y. Xu, Learning Discriminative Representations for Fine-Grained Diabetic Retinopathy GradingarXiv, 2011, p. 2120. URL, http://arxiv.org/abs/2011.02120.
[38] T. Araújo, G. Aresta, L. Mendonça, S. Penas, C. Maia, Â. Carneiro, A. M. Mendonça, A. Campilho, DR| GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images, Med. Image Anal. 63. arXiv:1910.11777, doi:10.1016/j.media.2020.101715.
[39] A. Singh, S. Sengupta, V. Lakshminarayanan, Explainable deep learning models in medical image analysis, J. Imaging 6 (6) (2020) 1–19, https://doi.org/10.3390/JIMAGING6060052, arXiv:2005.13799.
[40] S. Moccia, S.J. Wirkert, H. Kenngott, A.S. Vemuri, M. Apitz, B. Mayer, E. De Momi, L.S. Mattos, L. Maier-Hein, Uncertainty-aware organ classification for surgical data science applications in laparoscopy, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 65 (11) (2018) 2649–2659, https://doi.org/10.1109/TBME.2018.2813015, arXiv:1706.07002.
[41] T.J. Adler, L. Ardizzone, A. Vemuri, L. Ayala, J. Gröhl, T. Kirchner, S. Wirkert, J. Kruse, C. Rother, U. Köthe, L. Maier-Hein, Uncertainty-aware performance

assessment of optical imaging modalities with invertible neural networks, Int. J. Comput. Assist. Radiol. Surg. 14 (6) (2019) 997–1007, https://doi.org/10.1007/s11548-019-01939-9, arXiv:1903.03441.

[42] A. Kendall, Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision?, in: Advances in Neural Information Processing Systems 2017-December (Nips), 2017, pp. 5575–5585, arXiv:1703.04977.

[43] C. Leibig, V. Allken, M.S. Ayhan, P. Berens, S. Wahl, Leveraging uncertainty information from deep neural networks for disease detection, Sci. Rep. 7 (1) (2017) 1–14, https://doi.org/10.1038/s41598-017-17876-z.

[44] A. Rahimi, B. Recht, Random features for large-scale kernel machines, in: Advances in Neural Information Processing Systems 20 - Proceedings of the 2007 Conference, 2009.

[45] P.L. Gunawardhana, R. Jayathilake, Y. Withanage, G.U. Ganegoda, Automatic diagnosis of diabetic retinopathy using machine learning: a review, in: 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.

[46] S. Otálora, O. Perdomo, F. González, H. Müller, Training deep convolutional neural networks with active learning for exudate classification in eye fundus images, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10552, LNCS, 2017, pp. 146–154, https://doi.org/10.1007/978-3-319-67534-3_16. URL, http://www.who.int/diabetes/en/.

[47] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, arXiv: 026218253X, The MIT Press, 2006, https://doi.org/10.1142/S0129065704001899. URL, http://www.gaussianprocess.org/gpml/chapters/RW.pdf.

[48] K. Cutajar, E.V. Bonilla, P. Michiardi, M. Filippone, Random feature expansions for deep Gaussian processes, in: 34th International conference on machine learning, ICML 2017 2, 2017, pp. 1467–1482, arXiv:1610.04386.

[49] M.D. Abràmoff, J.C. Folk, D.P. Han, J.D. Walker, D.F. Williams, S.R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang, M. Lamard, D.C. Moga, G. Quellec, M. Niemeijer, Automated analysis of retinal images for detection of referable diabetic retinopathy, JAMA Ophthalmol. 131 (3) (2013) 351–357, https://doi.org/10.1001/jamaophthalmol.2013.1743.

[50] M. Voets, K. Møllersen, L.A. Bongo, 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) (2019) 1–11.

[51] B. Garg, N. Manwani, Robust deep ordinal regression under label noise, URL, in: S. J. Pan, M. Sugiyama (Eds.), Proceedings of the 12th Asian Conference on Machine Learning vol. 129, 2020, pp. 782–796. of Proceedings of Machine Learning Research, PMLR, Bangkok, Thailand, http://proceedings.mlr.press/v129/garg20a.html.

[52] J. Ren, I. Hacihaliloglu, E.A. Singer, D.J. Foran, X. Qi, Unsupervised domain adaptation for classification of histopathology whole-slide images, Front. Bioeng. Biotechnol. 7 (May) (2019) 1–12, https://doi.org/10.3389/fbioe.2019.00102.