

CNN explanation methods for ordinal regression tasks

Javier Barbero-Gómez^a, Ricardo P.M. Cruz^{b,c,*}, Jaime S. Cardoso^{b,c}, Pedro A. Gutiérrez^a, César Hervás-Martínez^a

^a Department of Computer Science and Numerical Analysis, University of Córdoba, Campus de Rabanales, Córdoba, 14071, Córdoba, Spain

^b Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias s/n, Porto, 4200-465, Porto, Portugal

^c INESC TEC, Rua Dr. Roberto Frias s/n, Porto, 4200-465, Porto, Portugal

ARTICLE INFO

Communicated by R. Mao

Keywords:

Convolutional neural networks
Explanation methods
Ordinal regression

ABSTRACT

The use of Convolutional Neural Network (CNN) models for image classification tasks has gained significant popularity. However, the lack of interpretability in CNN models poses challenges for debugging and validation. To address this issue, various explanation methods have been developed to provide insights into CNN models. This paper focuses on the validity of these explanation methods for ordinal regression tasks, where the classes have a predefined order relationship. Different modifications are proposed for two explanation methods to exploit the ordinal relationships between classes: Grad-CAM based on Ordinal Binary Decomposition (GradOBD-CAM) and Ordinal Information Bottleneck Analysis (OIBA). The performance of these modified methods is compared to existing popular alternatives. Experimental results demonstrate that GradOBD-CAM outperforms other methods in terms of interpretability for three out of four datasets, while OIBA achieves superior performance compared to IBA.

1. Introduction

The use of Convolutional Neural Network (CNN) models has become increasingly popular for image classification tasks such as object detection, face recognition, and medical diagnosis. Despite their success, one significant challenge is their lack of interpretability, which hinders their debugging, validation, and auditing. Researchers have developed various explanation methods to address this issue that provide insight into the inner workings of CNN models.

The objective of the most common type of explanation methods for CNN models is to generate an explanation map that highlights the (least and most) relevant regions of an image that were used by the model to make its classification decision. This map is represented by a matrix $E \in \mathbb{R}^{H \times W}$, where H and W are the height and width of the input image, respectively. The values in E range between 0 and 1, with 0 indicating no relevance and 1 indicating maximum relevance for each pixel in the original image.

These methods range from occlusion analysis [1] to sophisticated techniques such as backpropagation of gradients, which generate saliency maps [2,3]. On the other hand, Class Activation Maps (CAM) methods combine gradients with layer activations [4,5], with Grad-CAM as one of the most popular algorithms [5]. Finally, the Information Bottleneck for Attribution (IBA) method [6] is based on including a

perturbation in the network, which results in a bottleneck able to evaluate how important each region is for the final output. Note that the resulting explanation maps obtained by any of these methods are often used for object localization tasks [7]. However, there is a lack of research on the validity of these methods for ordinal regression tasks. Ordinal regression is a specific type of classification task where the discrete classes have a predefined order relationship.

In a traditional classification task, the goal is to assign a label $y \in \mathcal{Y}$ selected from a finite set of Q categories $\mathcal{Y} = \{C_1, C_2, \dots, C_Q\}$ to an input vector $\mathbf{x} \in \mathcal{X}$. A classification model is a mapping $g: \mathcal{X} \rightarrow \mathcal{Y}$ that predicts the labels of arbitrary patterns. The optimal parameters of said model can be set by using a training dataset $D = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}, i \in 1, \dots, N\}$.

In ordinal regression, the class labels have a specific ordering relation $<: C_1 < C_2 < \dots < C_Q$. This is similar to traditional regression tasks, where the output is a real value that can be ordered by the $<$ operator. However, in ordinal regression, the labels are discrete, and the task is to obtain a predicted label $\hat{y} \in \mathcal{Y}$ that is as close as possible to the actual label y regarding the ordering of the classes [8]. In this context, the set of relevant performance metrics differs from the traditional ones, as different types of errors have different costs. Examples of ordinal regression include age estimation [9] or biomedical tasks such as tumor level of malignancy [10], as explored later on.

* Correspondence to: Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias s/n, Porto, Portugal.

E-mail addresses: jbarbero@uco.es (J. Barbero-Gómez), rpacruz@fe.up.pt (R.P.M. Cruz), jaime.cardoso@fe.up.pt (J.S. Cardoso), pagutierrez@uco.es (P.A. Gutiérrez), chervas@uco.es (C. Hervás-Martínez).

<https://doi.org/10.1016/j.neucom.2024.128878>

Received 26 May 2024; Received in revised form 18 October 2024; Accepted 6 November 2024

Available online 17 November 2024

0925-2312/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

In a previous conference work [11], existing (nominal) explanation methods were evaluated on ordinal models, showing that the ordinal models were able to produce better explanations than a conventional (nominal) model optimized with cross-entropy. Experiments were performed across four different datasets, and the explanations were evaluated by occluding the least and most relevant areas given by the explanation. The statistical analysis showed that the ordinal methods not only outperformed cross-entropy in performance metrics (except sometimes for accuracy, which does not take ranking into consideration), but they also produced better explanations. This has motivated the present work, which devises specific ordinal explanation methods to exploit ordinality further and improve the explanations obtained.

The goals of this work can be summarized as follows:

- To propose the GradOBD-CAM and OIBA explanation methods as modifications to Grad-CAM and IBA, respectively, integrating ordinal information in producing the explanation map.
- To propose an evaluation procedure for explanation maps considering ordinal regression performance.
- To test these new methods using the novel evaluation procedure with four ordinal regression datasets.
- To evaluate the significance of these results using statistical tests.

This work is structured as follows: Firstly, Section 2 presents related work on ordinal regression and explainability. The proposed methods are elaborated in Section 3. The experimental protocol is laid out in Section 4 and respective results in Section 5. The work concludes with Section 6.

2. Related work

This section focuses on the underlying methods for the proposed method. While several ordinal regression methods exist, Section 2.1 focuses on Ordinal Binary Decomposition as the base method used in the experiments. Then, Sections 2.2–2.5 focus on four interpretability methods that are later explored (Grad-CAM, Grad-CAM++, Score-CAM, and IBA).

2.1. Ordinal binary decomposition

Losses such as cross-entropy or mean squared error are typically used in ordinal regression, albeit they assume, respectively, that the task is categorical (without taking advantage of the ordinality) or just traditional regression (which assumes a specific cardinality between the classes). An alternative technique for adapting CNNs to handle ordinal outputs was introduced in [12]. This approach involves breaking down the original classification task with Q classes into $Q - 1$ separate binary decision problems. It is commonly referred to as the Ordinal Binary Decomposition (OBD) method. In each of these binary problems, the objective is to determine whether the label y is ranked higher than C_q for a given sample x , where $1 \leq q < Q$ (as described in the “Ordered partitions” scheme in [8], which was firstly considered in [13]).

To implement this idea, the model’s output $\mathbf{o}(x) = (o_1, o_2, \dots, o_{Q-1})$ is designed with $Q - 1$ neurons using sigmoid activation. Each output o_q aims to predict the probability $P(y > C_q | x)$.

Fig. 1 shows the space of output \mathbf{o} probability possibilities. Notice that not all solutions are ordinally-consistent; for example, to predict weather (cold, warm, hot), $\mathbf{o} = (1, 0, 1)$ would be inconsistent. ECOC promotes the output distribution to follow a cumulative mass function by promoting the output to follow one of the dot points, \mathbf{v}_i in this specific case, C_2 .

As the OBD model produces cumulative probabilities, it requires combining these multiple outputs to make a decision. To address the issue of inconsistent probabilities, the approach employs ideal output vectors for each class q , namely $\mathbf{v}(C_q)$, based on the Error-Correcting Output Codes (ECOC) framework. The decision rule involves predicting

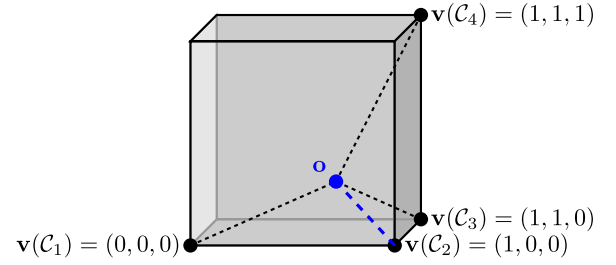


Fig. 1. Visualization of the model output vector \mathbf{o} (blue dot) for sample x and its distances to the ideal class vectors (dashed lines) in a 3D graphical representation. Each dimension corresponds to one of the three model outputs. The closest vector is $\mathbf{v}(C_2)$ (marked in blue), leading to the assignment of label C for sample x .

the class \hat{y}_i whose ideal vector minimizes the distance from the obtained output vector $\mathbf{o}(x)$ using the L_2 norm as the distance measure:

$$\mathbf{v}(C_q) = (v_1^q, v_2^q, \dots, v_{Q-1}^q), \quad (1)$$

$$v_c^q = \mathbb{1}\{c < q\}, \quad (2)$$

$$\hat{y}_i = \arg \min_{C_1 \leq C_q \leq C_Q} \|\mathbf{o}(x_i) - \mathbf{v}(C_q)\|_2, \quad (3)$$

where $\mathbb{1}\{\cdot\}$ is the indicator function that evaluates to 1 when the argument condition is true and 0 otherwise. Apart from the predicted class, a score can be assigned for each pattern x and class q , $s_q(x)$, based on the negative distance to the ideal vector:

$$f_q(x_i) = -\|\mathbf{o}(x_i) - \mathbf{v}(C_q)\|_2. \quad (4)$$

Finally, to adhere to this framework, the standard categorical cross-entropy loss is replaced with the squared error loss because it aligns better with the distance function used in the ECOC decision process:

$$\mathcal{L}_{SE}(x_i) = \sum_{q=1}^{Q-1} (\mathbb{1}\{y_i > C_q\} - P(y_i > C_q | x_i))^2. \quad (5)$$

2.2. Grad-CAM

The activation-based method Grad-CAM [5] aims to provide an explanation map that highlights the regions of the input image that are relevant for a CNN’s prediction by using the activations of the different features of an intermediate convolutional layer.

Let A_{ij}^k be the pixel in the i th row and j th column of feature map k of the selected intermediate layer A , which has height H_A and width W_A . A relevance score to each feature k , α_q^k , is computed to explain the output score of the target class f_q ; Grad-CAM then uses the average gradient with respect to the output:

$$\alpha_q^k = \frac{1}{H_A \times W_A} \sum_i \sum_j \frac{\partial f_q}{\partial A_{ij}^k}. \quad (6)$$

Finally, the explanation map is computed as a linear combination of the activation feature maps A^k based on their importance α_q^k , followed by the application of the Rectified Linear Unit (ReLU) to eliminate negative interactions:

$$E = \text{ReLU} \left(\sum_k \alpha_q^k A^k \right). \quad (7)$$

This explanation may then be upsampled to the same size as the input image and can be superimposed for easy visualization.

2.3. Grad-CAM++

Grad-CAM performs a weighted average across the feature maps but an unweighted average across the locations within each feature map. The authors of Grad-CAM++ [14] argue that computing the importance α_q^k of each feature map k as a global spatial average as in (6) makes

Grad-CAM less capable of properly localizing objects when multiple occurrences exist of objects of the same class, since one feature map may find the object in one region while another feature map may find the object in another region. Furthermore, the authors argue that the localization of the object may not correspond to the entire object but only a part of it.

To address this, they redefine α_q^k in the following manner:

$$\alpha_q^k = \sum_i \sum_j \gamma_{kij}^q \text{ReLU} \left(\frac{\partial f_q}{\partial A_{ij}^k} \right), \quad (8)$$

where a new weight γ_{kij}^q is introduced for each pixel in the activation that takes into account the second and third derivative information [14].

2.4. Score-CAM

In Score-CAM [15], the authors offer an alternative way to find α_q^k without using gradients. Gradients can be saturated or vanish due to the Sigmoid and ReLU non-linearities. Furthermore, the authors also argue that the unweighted average pooling is problematic.

The authors propose finding the importance of each feature map k as:

$$\alpha_q^k = f_q(\mathbf{x} \circ B^k) - f_q(\mathbf{x}_b), \quad (9)$$

where B^k is the upsampled and normalized version of the feature map A^k with dimensions $H \times W$ and all values ranging in the $[0, 1]$ interval, \circ is the element-wise product, and \mathbf{x}_b is a baseline image (for example, a black image).

The intuition is that the importance of each feature map should be measured by how much the restriction of the image to the regions highlighted by the activation map affects the output when compared to a neutral image.

2.5. IBA

The IBA method, as proposed by [6], is a perturbation-based method, meaning that some information is introduced in the computation to study its effects on the output of the model. However, unlike other perturbation methods that alter the information at the input of the model, it consists of injecting a perturbation amidst its information flow, creating a bottleneck in the network. This bottleneck helps evaluate the impact on the output of the regions from the input image.

To achieve this, it introduces a new random variable Z that maximizes the amount of information it shares with the output score of the target class $s_q(\mathbf{x})$ while minimizing the information it shares with the model input \mathbf{x} :

$$\max I[s_q(\mathbf{x}); Z] - \beta I[\mathbf{x}; Z], \quad (10)$$

where I denotes the mutual information, and β controls the trade-off between predicting the labels well and using little input information. $Z \in \mathbb{R}^{H_A \times W_A}$ acts as a substitute for the output of one of the intermediate layers A adding a certain noise $\epsilon \in \mathbb{R}^{H_A \times W_A}$:

$$Z = \lambda(\mathbf{x})A + (1 - \lambda(\mathbf{x}))\epsilon, \quad (11)$$

where $\lambda(\mathbf{x}) \in [0, 1]^{H_A \times W_A}$ adjusts how much of the original signal is passed along (see Fig. 2).

To obtain a value of $\lambda(\mathbf{x})$ that aligns with the objective posed in Eq. (10), a loss function \mathcal{L}_λ is designed. To estimate how much information from A is passed along in Z , mutual information is used:

$$I[A; Z] = \mathbb{E}_A[D_{\text{KL}}[P(Z | A) \| P(Z)]], \quad (12)$$

where $P(Z | A)$ and $P(Z)$ are the respective probability distributions, D_{KL} is the Kullback–Leibler divergence and \mathbb{E}_A the expectation over A . This, however, is an unmanageable computation, so an approximation $Q(Z) = \mathcal{N}(\mu_A, \sigma_A)$ is made assuming that all dimensions of Z are

distributed independently and normally, which overestimates the real value:

$$I[A; Z] = \mathbb{E}_A[D_{\text{KL}}[P(Z | A) \| Q(Z)]] - D_{\text{KL}}[P(Z) \| Q(Z)] \quad (13)$$

Finally, the information loss function \mathcal{L}_I is defined as:

$$\mathcal{L}_I = \mathbb{E}_A[D_{\text{KL}}[P(Z | A) \| Q(Z)]], \quad (14)$$

and the final loss function \mathcal{L}_λ is defined as the combination of \mathcal{L}_I and the cross-entropy loss \mathcal{L}_{CE} :

$$\mathcal{L}_\lambda = \mathcal{L}_{\text{CE}} + \beta \mathcal{L}_I. \quad (15)$$

This can now be used to optimize $\lambda(\mathbf{x})$, parameterized as $\lambda(\mathbf{x}) = \sigma(\alpha(\mathbf{x}))$ (where $\alpha \in \mathbb{R}^{H_A \times W_A}$ and σ is the sigmoid function), by minimizing \mathcal{L}_λ using any stochastic gradient descent algorithm such as the Adam backpropagation method [16].

Regions of the image with relevant information will present a λ value close to 1; conversely, irrelevant parts will present a value close to 0. For this reason, the output explanation map E is just λ upsampled to the original input size.

3. Methodology

While the validity and performance of the methods shown in the previous section have been evaluated for traditional nominal classification tasks, neither Grad-CAM nor IBA considers the order relationship between the class labels in ordinal regression tasks.

In this section, we propose two new explanation methods for CNN OBD models, based on both Grad-CAM and IBA, respectively, that use order information in their process to improve the ordinal performance of the resulting explanation map.

3.1. GradOBD-CAM

We propose an explanation method based on Grad-CAM, which uses gradient information about the activation of all output score neurons o_c in the OBD model to compute the feature coefficients α_q^k .

Recall that in an OBD model for Q -class ordinal classification, there are $Q - 1$ output neurons, each one corresponding to successive class threshold probabilities: $P(y > C_1)$, $P(y > C_2)$, ..., $P(y > C_{Q-1})$. For an input sample \mathbf{x} with class label y , more importance should be given to feature maps that contribute positively to the output probabilities $P(y > C_p | \mathbf{x})$ such that $y > C_p$ and, conversely, less importance should be given to feature maps that contribute negatively to the output probabilities $P(y > C_p | \mathbf{x})$ such that $y \leq C_p$. To this end, we introduce a new parameter δ_c^q :

$$\alpha_q^k = \sum_{c=1}^{Q-1} \delta_c^q \frac{1}{W_A \times H_A} \sum_i \sum_j \frac{\partial o_c}{\partial A_{ij}^k}, \quad (16)$$

such that

$$\delta_c^q = \begin{cases} +1 & \text{if } c < q, \\ -1 & \text{if } c \geq q. \end{cases} \quad (17)$$

The previous Grad-CAM equation [Eq. (6)] produced class-independent scores for each feature k is now made rank-dependent in the new score coefficient α_q^k so that more importance is given to feature maps that contribute positively to the output probabilities, and the converse as well. This should more naturally conform to the ordinality of the output and the OBD approach more specifically.

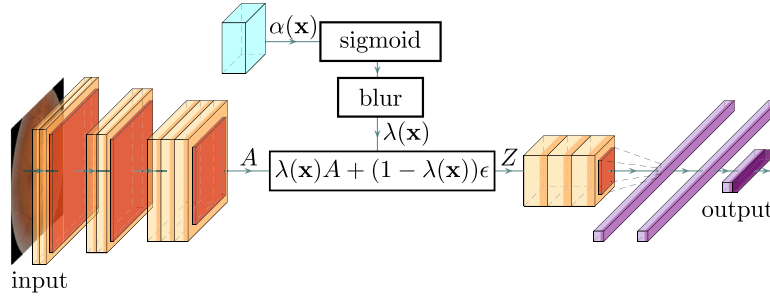


Fig. 2. IBA is a perturbation method which adds an information bottleneck in the middle of the network, where λ adjusts how much information is allowed to pass.

3.2. OIBA

Furthermore, we propose a modification to the perturbation-based method IBA that leverages the ordinal loss of the OBD model in the optimization process for $E = \lambda(\mathbf{x})$. We call this method the Ordinal IBA, or OIBA.

The information bottleneck is already covered by the information loss \mathcal{L}_I defined in Eq. (14), but the cross-entropy loss \mathcal{L}_{CE} is not well suited to the ordinal case. This component of the loss function can be substituted by the squared error loss defined in Eq. (5):

$$\mathcal{L}_\lambda = \mathcal{L}_{SE} + \beta \mathcal{L}_I. \quad (18)$$

This modification (the inclusion of \mathcal{L}_{SE}) introduces the same ranking prior of the OBD method relative to CE and allows the construction process of the explanation map to exploit information from all outputs of the model in its native representation.

3.3. Evaluating the performance of explanation methods

A commonly used technique is perturbation analysis to assess the impact of a specific region in an image on the classification decision. This involves occluding parts of the input images and observing the resulting model output changes. If the occluded regions, according to an explanation denoted as E_i , are deemed relevant, the classification performance should exhibit a specific pattern of decline.

An approach described in [14] suggests a simple method for implementing this idea. It involves multiplying the input \mathbf{x}_i by the explanation E_i to obtain an occluded image $\tilde{\mathbf{x}}_i$:

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i \circ E_i, \quad (19)$$

where \circ denotes element-wise multiplication. By examining the average effect on the score of the target class C_q , denoted as $f_q(\mathbf{x})$, for each example in which the score decreases, we can calculate the average drop:

$$\text{Average drop} = \frac{1}{N} \sum_i \frac{\max(0, f_q(\mathbf{x}_i) - f_q(\tilde{\mathbf{x}}_i))}{f_q(\mathbf{x}_i)}. \quad (20)$$

The objective is to minimize the average drop, as a good explanation should result in a minimal reduction in score. However, this metric does not account for the ordering information among class labels. In other words, when the confidence in the target class diminishes, the confidence in nearby classes should increase instead of in distant ones. Unfortunately, this aspect is not evaluated by the average drop metric.

An alternative approach, proposed by [6], involves dividing the explanation map into tiles, such as 8×8 tiles, and ranking them based on the total sum of relevance within each tile. The input image is then occluded tile by tile, starting from the most relevant tile and progressing to the least relevant. This process generates the Most Relevant Features (MoRF) curve, which plots the target class score against the level of image degradation (i.e., the number of occluded tiles). For a meaningful explanation map, the score is expected to decrease sharply at the beginning when the most relevant parts of the image are

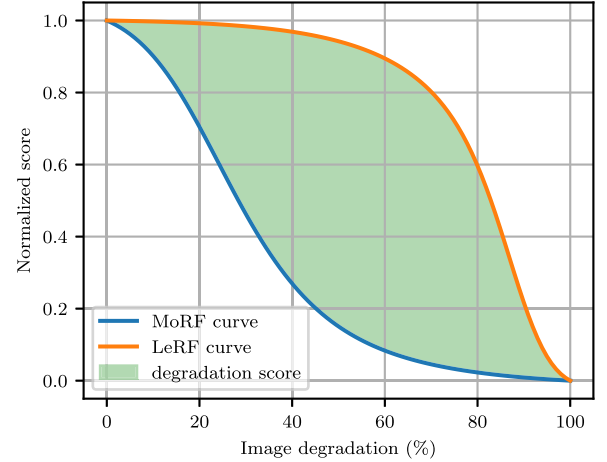


Fig. 3. The degradation score represents how much the predictive score degrades as the most relevant features (MoRF) and the least relevant features (LeRF) are occluded. Example of the MoRF and LeRF curves and the area between them or “degradation score”.

occluded. The same procedure is repeated in reverse order of relevance to obtain the Least Relevant Features (LeRF) curve. In this case, the score should not drop significantly until the most relevant parts are occluded. The extremes of these curves can be normalized between 0 and 1, and the signed area between them is computed. A relevant explanation map is expected to yield a large area between the MoRF and LeRF curves. An illustrative example is depicted in Fig. 3.

This approach offers an advantage: it enables the study of the behavior of any metric, not limited to the target score. Thus, we propose examining the degradation of the following classification performance metrics, most of them specific to ordinal regression [17]:

- Correct Classification Rate (CCR): the proportion of patterns that are assigned their correct label:

$$\text{CCR} = \frac{1}{N} \sum_{i=1}^N \mathbb{1} \{ \hat{y}_i = y_i \}, \quad (21)$$



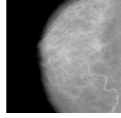
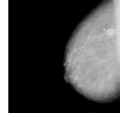
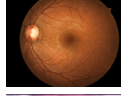


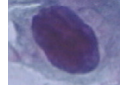
where CCR values range from 0 to 1 and should be maximized.

- Average Area under the ROC curve (AvAUC): the receiver operating characteristic (ROC) curve is computed for each class as a binary one-vs-rest problem and its area is computed. The average area over all the classes is taken to account for class balancing, it varies between 0 and 1 and should be maximized.
- Mean Absolute Error (MAE): an integer rank is assigned to each class $r(C_j) = j$ corresponding to their ordering. The MAE is the average absolute difference between the real and predicted rank:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |r(\hat{y}_i) - r(y_i)|, \quad (22)$$

where $\text{MAE} \in [0, Q - 1]$, and the metric should be minimized.

Table 1
Datasets used for the experiments.

Dataset	Number of observations	Number of classes	Illustration first class	Illustration last class
Adience [9]	17 702	8		
CBIS-DDSM [21]	2620	6		
Retinopathy ¹	53 569	5		
Herlev Pap-Smear [10]	917	4		

¹ <https://www.kaggle.com/c/diabetic-retinopathy-detection>

- Quadratic Weighted Kappa (κ): a measure of the agreement between the real and predicted labels [18]. Let w_{ij} be the disagreement cost when $y = C_i$ and $\hat{y} = C_j$ ($w_{ij} = (i - j)^2$), p_{ij} is the observed agreement and e_{ij} is the expected disagreement due to chance. κ is defined as:

$$\kappa = 1 - \frac{\sum_{i=1}^Q \sum_{j=1}^Q w_{ij} p_{ij}}{\sum_{i=1}^Q \sum_{j=1}^Q w_{ij} e_{ij}}, \quad (23)$$

with $\kappa \in [0, 1]$, and it should be maximized.

- Spearman's rank correlation coefficient (r_s): a non-parametric measure of rank correlation. If σ_y and $\sigma_{\hat{y}}$ are the standard deviations of the real and predicted labels respectively, and $\text{Cov}(y, \hat{y})$ is the covariance between the two variables, r_s is defined as:

$$r_s = \frac{\text{Cov}(y, \hat{y})}{\sigma_y \sigma_{\hat{y}}}, \quad (24)$$

where $r_s \in [0, 1]$, and the metric should be maximized.

4. Experiment design

4.1. Model and training

A ResNet34 CNN model [19] with OBD output pre-trained on ImageNet-1K [20] is used as the baseline model in all experiments. Training is done in batches of 64 samples for a maximum of 200 epochs, stopping early when the loss on the validation set does not improve for 20 epochs. This is repeated using a different random initialization and a different 80/10/10 training/validation/testing split 60 different times using the Adam backpropagation method [16]. Six different explanation methods are then applied: Grad-CAM [5], Grad-CAM++ [14], Score-CAM [15], IBA [6] and our two proposals, GradOBD-CAM and OIBA. Degradation of the following metrics is obtained as outlined in Section 3.3: CCR, AvAUC, MAE, κ and r_s . Notice that a higher degradation score is always better, independent of the underlying metric.

4.2. Data

Four image datasets are selected to test the performance of the explanation methods. A summary is presented in Table 1. They were chosen because they all present an ordinal prediction label and cover a large spectrum of different classification tasks with heterogeneous forms of images (facial photographs, cell cytology, eye exams, and mammograms). In all cases, the original images were resized to a resolution of 224×224 and normalized to ImageNet's original mean and standard deviation per input channel.

Adience dataset. A set of 17 702 photos of people scraped from the web and pre-aligned to fit their face, which comes already categorized into 8 different age groups [9] of increasing value: 0 to 2 years, 4 to 6 years, 8 to 13 years, 15 to 20 years, 25 to 32 years, 38 to 43 years, 48 to 53 years, and 60 years and up.

Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM). A database of 2620 scanned film mammography studies curated from the larger DDSM and each one assigned a Breast Imaging Reporting and Data System (BI-RADS) assessment [21] by a trained mammographer. The assessment is done on a scale of 0 to 5 according to the standard for a total of 6 classes (there are no cases of class 6 as there is no biopsy information). For this dataset, before being resized, all images were cropped into a square centered around the Region of Interest of the lesion.

Diabetic retinopathy dataset. A collection of 53 569 high-resolution retina images,¹ rated by a clinician on the presence of Diabetic Retinopathy (DR), an eye disease present in a large proportion of diabetes patients, on a scale from 0 (no DR) to 4 (proliferative DR) for a total of 5 classes.

Herlev pap smear dataset. 917 images of single Pap smear cells classified by doctors and technicians into 7 different classes, 3 of them normal from different parts of the cervix (242 images in total) and 4 of them abnormal in different stages of dysplasia (675 images in total) [10]. These are condensed into 4 ordinal classes, following the Bethesda System standard [22].

5. Results

In Figs. 4 and 5, boxplots for each metric degradation and dataset are shown. Some examples of the explanation maps are shown in Fig. 6.

GradOBD-CAM performance is generally better than Grad-CAM++ and Score-CAM and greater than or on par with Grad-CAM. The greatest difference in performance is observed in the Adience dataset, but it is also noticeable in others. In the case of the IBA methods, OIBA achieves a greater performance in all metrics for the Adience and Retinopathy datasets, although no observable difference is seen for the CBIS-DDSM and Herlev datasets.

Hypothesis testing is now performed to validate the results statistically. A Wilcoxon signed-rank test is performed for each metric and

¹ <https://www.kaggle.com/c/diabetic-retinopathy-detection/data>

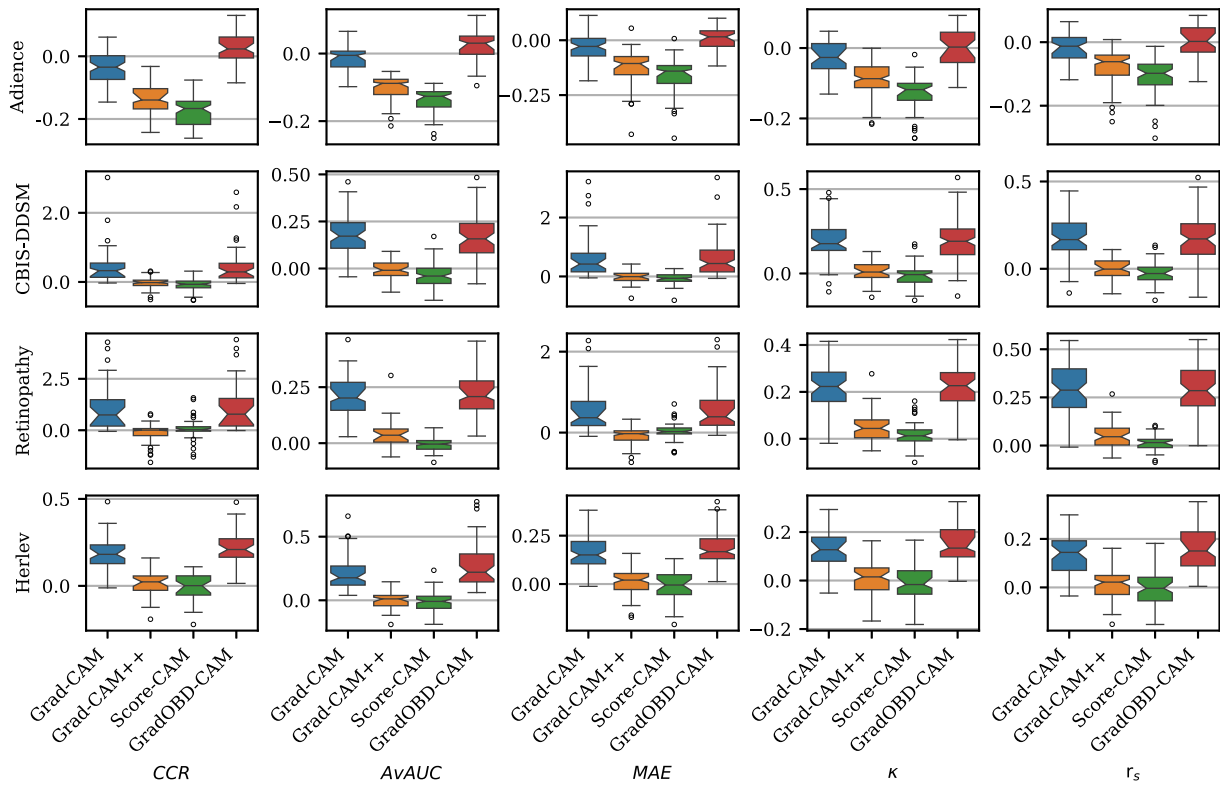


Fig. 4. Metric degradation boxplots of the CAM methods for each metric and dataset (higher is better). The box notches represent the confidence interval around the median. Points outside the whiskers represent outliers.

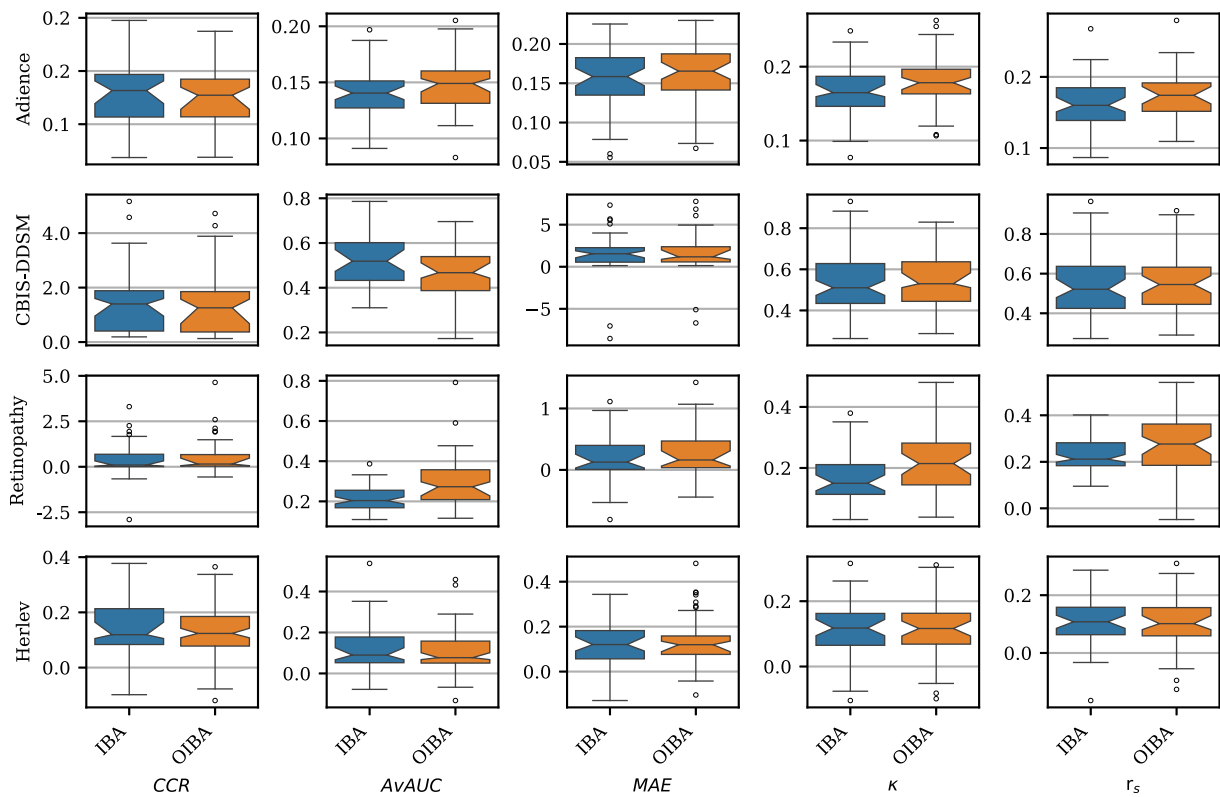


Fig. 5. Metric degradation boxplots of the IBA methods for each metric and dataset (higher is better). The box notches represent the confidence interval around the median. Points outside the whiskers represent outliers.

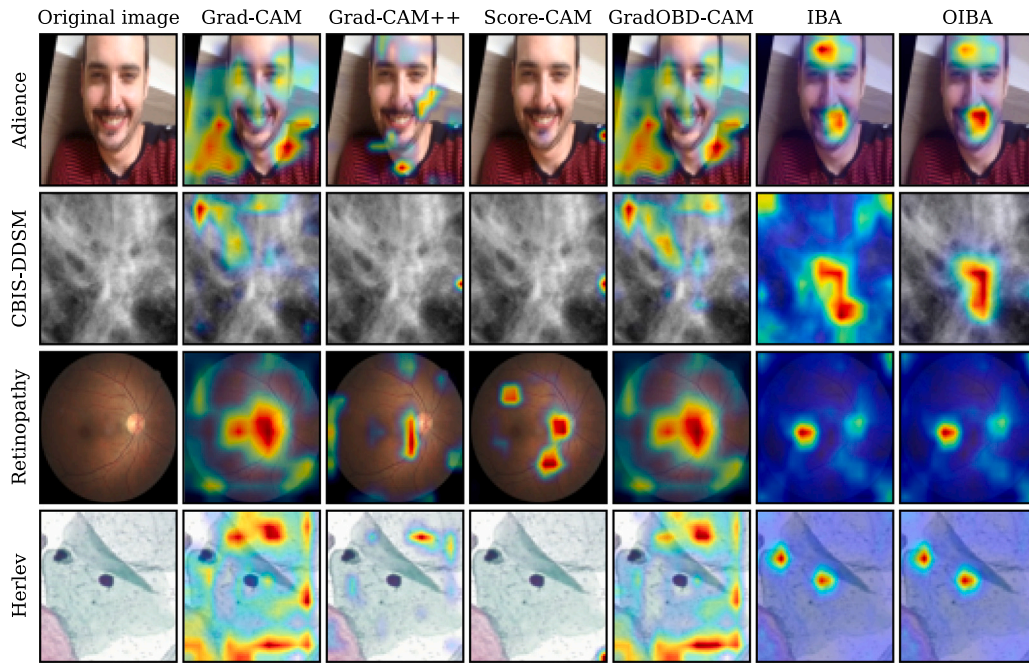


Fig. 6. Example explanation maps from each dataset generated by each one of the compared explanation methods.

Table 2

Wilcoxon signed-rank tests for the all datasets. Values less than $\alpha = 0.05$ are highlighted in **bold**.

	Grad-CAM vs GradOBD-CAM		Grad-CAM++ vs GradOBD-CAM		Score-CAM vs GradOBD-CAM		IBA vs OIBA	
	Z	p-value	Z	p-value	Z	p-value	Z	p-value
Adience								
CCR	-6.633	<.001	-6.736	<.001	-6.736	<.001	-5.168	<.001
AvAUC	-6.044	<.001	-6.736	<.001	-6.736	<.001	-5.241	<.001
MAE	-6.125	<.001	-6.633	<.001	-6.721	<.001	-6.000	<.001
κ	-5.713	<.001	-6.559	<.001	-6.706	<.001	-6.493	<.001
r_s	-5.536	<.001	-6.331	<.001	-6.655	<.001	-6.140	<.001
CBIS-DDSM								
CCR	-0.331	.740	-6.596	<.001	-6.633	<.001	-2.194	.028
AvAUC	-1.045	.296	-6.633	<.001	-6.677	<.001	-4.940	<.001
MAE	-3.143	.002	-6.589	<.001	-6.618	<.001	-1.561	.119
κ	-0.677	.498	-6.449	<.001	-6.434	<.001	-1.627	.104
r_s	-0.169	.866	-6.279	<.001	-6.383	<.001	-1.104	.269
Retinopathy								
CCR	-4.491	<.001	-6.721	<.001	-6.265	<.001	-3.946	<.001
AvAUC	-3.924	<.001	-6.714	<.001	-6.721	<.001	-6.221	<.001
MAE	-5.558	<.001	-6.699	<.001	-6.235	<.001	-5.661	<.001
κ	-4.881	<.001	-6.684	<.001	-6.684	<.001	-6.154	<.001
r_s	-0.044	.965	-6.721	<.001	-6.706	<.001	-3.092	.002
Herlev								
CCR	-4.454	<.001	-6.736	<.001	-6.736	<.001	-0.582	.561
AvAUC	-4.829	<.001	-6.736	<.001	-6.736	<.001	-1.524	.128
MAE	-4.807	<.001	-6.721	<.001	-6.736	<.001	-0.169	.866
κ	-3.997	<.001	-6.574	<.001	-6.684	<.001	-0.169	.866
r_s	-4.211	<.001	-6.603	<.001	-6.706	<.001	-0.670	.503

dataset to check whether the median performance of GradOBD-CAM is worse than or equal to the other three CAM methods (null hypothesis) or is, in fact, better (alternative hypothesis). The same is also performed to compare OIBA with IBA. A standard significance level of $\alpha = 0.05$ is used in all cases. The results of these tests can be seen in Table 2.

The tests show that GradOBD-CAM has a better median performance than both Grad-CAM++ and Score-CAM in all metrics for the Adience and Herlev datasets. For the CBIS-DDSM dataset, GradOBD-CAM performs better than Grad-CAM++ and Score-CAM in all metrics, but performs similarly to the original Grad-CAM. Finally, for the Retinopathy

dataset, GradOBD-CAM obtains a better median performance, except for r_s against Grad-CAM.

In the case of OIBA, it achieves greater median performance for two (Adience and Retinopathy) of the four tested datasets. In all other cases, performance does not drop for any metric.

The reason why Adience and Diabetic Retinopathy show gains more clearly may be due to the largest number of ordinal classes and subtleties in the dimension of complexity of these datasets. Smaller datasets like Herlev (with only 4 classes) may not highlight significant differences due to limited data variability.

6. Conclusions and future work

Two explanation methods, GradOBD-CAM and OIBA are proposed in this work. GradOBD-CAM adapts the Grad-CAM attribution method for the ordinal regression context using the OBD model. OIBA introduces the ordinal loss function native to the OBD model in the explanation process. Results for four datasets, measured using an interpretability metric based on the degradation score of ordinal metrics as the regions deemed most/least relevant are occluded, show that GradOBD-CAM was superior to all its CAM counterparts, at a statistically significant level, for three out of four datasets, with one exception for the Spearman's correlation coefficient. In the remaining dataset, the method was still statistically superior to the two interpretability methods. OIBA also performs superior to IBA for two datasets in all metrics.

In future work, we would like to design other improvements for interpretability methods that can be used with other ordinal regression models — such as parametric models where the neural network outputs a Binomial distribution [23] or loss-based methods that promote ordinality in the output probability distribution [24]. Furthermore, while ordinality is generally promoted on the output, it could be possible to take advantage of GradOBD-CAM to promote ordinality within the latent feature space. Since explanations are based on the latent space of the neural network, it could help to promote this space to be itself ordinal so that perturbations of this space would affect the probability distribution around the target class. A loss term could be introduced to penalize non-monotonic growth for each activation value across the ordinal classes in a batch of images. Finally, gradient-based methods were chosen for this work since they are more malleable and usually perform well. It would be interesting to consider how ordinality could be added to other families of methods, such as DeepLift, SHAP, or LRP.

CRedit authorship contribution statement

Javier Barbero-Gómez: Writing – original draft, Visualization, Validation, Methodology, Investigation. **Ricardo P.M. Cruz:** Writing – original draft, Methodology, Investigation. **Jaime S. Cardoso:** Writing – review & editing, Supervision, Conceptualization. **Pedro A. Gutiérrez:** Writing – review & editing, Supervision. **César Hervás-Martínez:** Writing – review & editing, Validation, Formal analysis.

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 has been partially subsidized by the “Agencia Estatal de Investigación” (Spain) [grant references: PID2020-115454GB-C22/AEI/10.13039/501100011033 and PID2023-150663NB-C22], by the European Commission, project Test and Experiment Facilities for the Agri-Food Domain, AgriFoodTEF [grant reference: DI-GI-TAL-2022-CLOUD-AI-02, 101100622], and by the ENIA International Chair in Agriculture, University of Córdoba [grant reference TSI-100921-2023-3], funded by the Secretary of State for Digitalisation and Artificial Intelligence and by the European Union - Next Generation EU. Javier Barbero-Gómez research has been subsidized by the FPI Predoctoral Program of the “Ministerio de Ciencia, Innovación y Universidades” (Spain) [grant reference PRE2018-085659].

References

- [1] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13, Springer, 2014, pp. 818–833.
- [2] K. Simonyan, A. Vedaldi, A. Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, Tech. Rep., 2014, arXiv:1312.6034, arXiv.
- [3] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for Simplicity: The All Convolutional Net, Tech. Rep., 2015, arXiv:1412.6806, arXiv.
- [4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning Deep Features for Discriminative Localization, Tech. Rep., 2015, arXiv:1512.04150, arXiv.
- [5] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
- [6] K. Schulz, L. Sixt, F. Tombari, T. Landgraf, Restricting the Flow: Information Bottlenecks for Attribution, Tech. Rep., 2020, arXiv:2001.00396, arXiv.
- [7] W. Hui, C. Tan, G. Gu, Y. Zhao, Gradient-based refined class activation map for weakly supervised object localization, Pattern Recognit. 128 (2022) 108664, <http://dx.doi.org/10.1016/j.patcog.2022.108664>.
- [8] 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, <http://dx.doi.org/10.1109/TKDE.2015.2457911>.
- [9] E. Eidinger, R. Enbar, T. Hassner, Age and gender estimation of unfiltered faces, IEEE Trans. Inf. Forensics Secur. 9 (12) (2014) 2170–2179, <http://dx.doi.org/10.1109/TIFS.2014.2359646>.
- [10] J. Jantzen, J. Norup, G. Dounias, B. Bjerregaard, Pap-smear benchmark data for pattern classification, 2005, Nature Inspired Smart Information Systems (NiSIS).
- [11] J. Barbero-Gómez, R. Cruz, J.S. Cardoso, P.A. Gutiérrez, C. Hervás-Martínez, Evaluating the performance of explanation methods on ordinal regression CNN models, in: International Work-Conference on Artificial Neural Networks, IWANN 2023, Ponta Delgada, Portugal, 2023, pp. 529–540.
- [12] J. Barbero-Gómez, P.A. Gutiérrez, C. Hervás-Martínez, Error-correcting output codes in the framework of deep ordinal classification, Neural Process. Lett. (2022) <http://dx.doi.org/10.1007/s11063-022-10824-7>.
- [13] E. Frank, M. Hall, A simple approach to ordinal classification, in: European Conference on Machine Learning, Springer, Berlin, Heidelberg, Freiburg, Germany, 2001, pp. 145–156, http://dx.doi.org/10.1007/3-540-44795-4_13.
- [14] A. Chattopadhyay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM++: Improved visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conf. Appl. Comput. Vis., WACV, 2018, pp. 839–847, <http://dx.doi.org/10.1109/WACV.2018.00097>.
- [15] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score-CAM: Score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 24–25, <http://dx.doi.org/10.1109/CVPRW50498.2020.00020>.
- [16] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, Tech. Rep., 2017, arXiv:1412.6980, arXiv.
- [17] M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero, P.A. Gutiérrez, Metrics to guide a multi-objective evolutionary algorithm for ordinal classification, Neurocomputing 135 (2014) 21–31, <http://dx.doi.org/10.1016/j.neucom.2013.05.058>.
- [18] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46, <http://dx.doi.org/10.1177/001316446002000104>.
- [19] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
- [20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255, <http://dx.doi.org/10.1109/CVPR.2009.5206848>.
- [21] R.S. Lee, F. Gimenez, A. Hoogi, K.K. Miyake, M. Gorovoy, D.L. Rubin, A curated mammography data set for use in computer-aided detection and diagnosis research, Sci. Data 4 (1) (2017) 1–9, <http://dx.doi.org/10.1038/sdata.2017.177>.
- [22] R. Nayar, D.C. Wilbur, The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes, third ed., Springer, 2015.
- [23] J.P. Costa, J.S. Cardoso, Classification of ordinal data using neural networks, in: Machine Learning: ECML 2005: 16th European Conference on Machine Learning, Porto, Portugal, October 3–7, 2005. Proceedings 16, Springer, 2005, pp. 690–697.
- [24] T. Albuquerque, R. Cruz, J.S. Cardoso, Ordinal losses for classification of cervical cancer risk, PeerJ Comput. Sci. 7 (2021) e457, <http://dx.doi.org/10.7717/peerj-cs.457>.



Javier Barbero-Gómez received his B.S. degree in Computer Engineering in 2017 and his M.S. degree in Computer Engineering in 2019 from the University of Córdoba (Spain). He has earned his Ph.D. in 2024 on Deep Learning for ordinal multi-dimensional data. He is a member of the Computer Science and Numerical Analysis Department of the University of Córdoba. His research interests include Deep Learning, ordinal classification and applications such as medical imaging.



Ricardo P.M. Cruz received a B.S. degree in Computer Science and an M.S. degree in Applied Mathematics, both from the University of Porto, Portugal. Since 2015, he has been a researcher at INESC TEC, working in machine learning with a particular emphasis on computer vision. He earned his Ph.D. in Computer Science in 2021, specializing in computer vision and deep learning. Currently, he is a post-doctoral researcher on autonomous driving under the THEIA research project, a partnership between the University of Porto and Bosch Car Multimedia.



Jaime S. Cardoso received his B.S. degree in Electrical and Computer Engineering in 1999, his M.Sc. in Mathematical Engineering in 2005, and his Ph.D. in Computer Vision in 2006, all from the University of Porto. Cardoso is a Full Professor at the Faculty of Engineering of the University of Porto and a Researcher at INESC TEC. From 2012 to 2015, Cardoso served as President of the Portuguese Association for Pattern Recognition, affiliated with the IAPR. Cardoso is also a Senior Member of IEEE since 2011. Cardoso



Pedro A. Gutiérrez received his B.S. degree in Computer Science from the University of Sevilla (Spain) in 2006, and his Ph.D. degree in Computer Science and Artificial Intelligence from the University of Granada (Spain) in 2009. He is currently a Professor with the Department of Computer Science and Numerical Analysis, University of Córdoba (Spain). His research interests are in the areas of supervised learning, evolutionary artificial neural networks, ordinal classification. He serves on the Editorial board for the journals IEEE Transaction on Neural Networks and Learning Systems and on the organisation/program committees of several computational intelligence conferences.



César Hervás-Martínez received his B.S. degree in Statistics and Operations Research from the Universidad Complutense de Madrid (Spain) in 1978, and his Ph.D. degree in Mathematics from the University of Seville (Spain) in 1986. He is currently a Professor of Computer Science and Artificial Intelligence with the Department of Computer Science and Numerical Analysis, University of Córdoba (Spain). His current research interests include neural networks, evolutionary computation, and the modeling of natural systems.