

## Article

# Quasi-Unimodal Distributions for Ordinal Classification

Tomé Albuquerque <sup>1,2,\*</sup> , Ricardo Cruz <sup>1,2</sup> and Jaime S. Cardoso <sup>1,2</sup> 
<sup>1</sup> Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal; rpcruz@fe.up.pt (R.C.); jaime.cardoso@inesctec.pt (J.S.C.)

<sup>2</sup> Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal

\* Correspondence: tome.m.albuquerque@inesctec.pt; Tel.: +351-91-277-2944

**Abstract:** Ordinal classification tasks are present in a large number of different domains. However, common losses for deep neural networks, such as cross-entropy, do not properly weight the relative ordering between classes. For that reason, many losses have been proposed in the literature, which model the output probabilities as following a unimodal distribution. This manuscript reviews many of these losses on three different datasets and suggests a potential improvement that focuses the unimodal constraint on the neighborhood around the true class, allowing for a more flexible distribution, aptly called quasi-unimodal loss. For this purpose, two constraints are proposed: A first constraint concerns the relative order of the top-three probabilities, and a second constraint ensures that the remaining output probabilities are not higher than the top three. Therefore, gradient descent focuses on improving the decision boundary around the true class in detriment to the more distant classes. The proposed loss is found to be competitive in several cases.

**Keywords:** ordinal classification; ordinal losses; deep learning; optimization



**Citation:** Albuquerque, T.; Cruz, R.; Cardoso, J.S. Quasi-Unimodal Distributions for Ordinal Classification. *Mathematics* **2022**, *10*, 980. <https://doi.org/10.3390/math10060980>

Academic Editor: Paolo Crippa

Received: 31 December 2021

Accepted: 16 March 2022

Published: 18 March 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

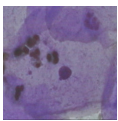

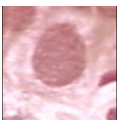
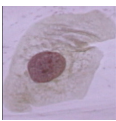
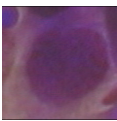
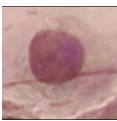
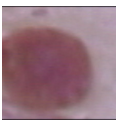


**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Since the emergence of deep learning models, cross-entropy has been the most used loss function due to its simplicity, differentiability, and competence to deal with many classes. When a model is trained with cross-entropy there is no prior knowledge of the class labels, which means that all semantic relationships that might exist between the labels are ignored. However, several classification tasks expose a natural order/structure between the labels, which means that the labels present a specific order which may be useful as prior knowledge for the deep models [1]. Cancer grading is a typical example of an ordinal problem, wherein images show a natural progression from normal cells undergoing a gradual process to cancer cells, presenting different stages between the two terminal stages (normal and cancer) [2] (see Table 1).

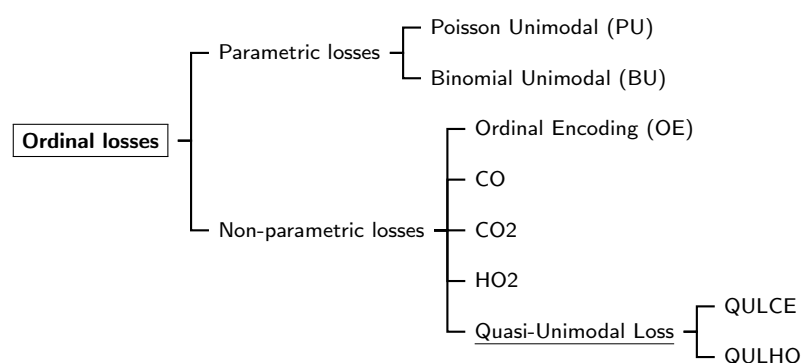
**Table 1.** Illustration with images of the seven different classes in the Herlev dataset.

Normal				Abnormal			
							
WHO $k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	
TBS	$k = 1$	$k = 2$	$k = 3$	$k = 4$			

While cross-entropy focuses on each output probability in isolation, ordinal classification focuses on the relative order of the output probabilities, the rationale being that the

error between two neighbor classes should be lower than the error between two distant classes. This makes ordinal classification more suitable in a variety of situations such as age estimation [3–5], movies rating [6,7], cancer grading [2,8], market bonds rating [9], emotion estimation [10–12], photographs dating [13], diabetic retinopathy grading [14] and gene expression analysis [15], in addition to many others.

A large body of work in the literature proposes losses that promote a unimodal distribution in class labels  $p(y|x)$ . This is often achieved with one of the two main classes of ordinal losses (as illustrated by Figure 1): Parametric losses and non-parametric losses. Parametric losses try to antecedently impose unimodality on the posterior distributions using a single penalty on all the labels. This is done, for example, by assuming that the output function follows a Poisson distribution (PU) [14] or a binomial distribution (BU) [16]. Regarding non-parametric losses, among several in the literature, Ordinal Encoding (OE) [17] or HO2 [2] try not to confine the learned representation to a single parametric model, so that a wider range of possible output functions can be explored, avoiding ad hoc decisions. All the losses mentioned above will be explored in more detail in the next section.



**Figure 1.** Schematic representation of ordinal loss functions from the literature. The proposed loss function is underlined.

**Contributions:** For neural networks, a novel non-parametric ordinal loss is presented that induces output probabilities to follow a quasi-unimodal distribution. This is accomplished by applying two different constraints: The first restriction is between the neighboring pair of output probabilities (relative to the true class). Then, a second constrain penalizes all remaining probabilities above that neighborhood, offering a more flexible decision boundary.

## 2. Related Work

Assume that in a classification task, the instances have one of  $K$  classes, whose labels are  $\mathcal{C}^{(1)}$  to  $\mathcal{C}^{(K)}$ , which correspond to the natural order of the ordinal classes.

**Cross-Entropy (CE):** Typically, a neural network is trained to perform multi-class classification by minimizing the cross-entropy loss for the entire training set,

$$\text{CE}(\mathbf{y}_n, \hat{\mathbf{y}}_n) = - \sum_{k=1}^K y_{nk} \log(\hat{y}_{nk}), \quad (1)$$

where, for each  $n$ -th observation,  $\mathbf{y}_n = [y_{n1} \cdots y_{nk} \cdots y_{nK}] \in R^K$ , with  $y_{nk} \in \{0, 1\}$ , representing its respective one-hot encoding and  $\hat{\mathbf{y}}_n = [\hat{y}_{n1} \cdots \hat{y}_{nk} \cdots \hat{y}_{nK}] \in R^K$ , with  $\hat{y}_{nk} \in [0, 1]$  being the respective vector of output probabilities assigned by the model for that  $n$ -th observation. Naturally,  $\sum_{k=1}^K y_{nk} = \sum_{k=1}^K \hat{y}_{nk} = 1$ .

However, CE has limitations when applied to ordinal data. Defining  $k_n^* \in \{1, \dots, K\}$  as the index of the true class of observation  $\mathbf{x}_n$  (the position where  $y_{nk} = 1$ ), it is then clear that

$$\text{CE}(\mathbf{y}_n, \hat{\mathbf{y}}_n) = - \log(\hat{y}_{nk_n^*}). \quad (2)$$

Intuitively, CE is just trying to maximize the probability in the output corresponding to the true class, ignoring all the other probabilities. For this loss, an error between classes  $\mathcal{C}^{(1)}$  and  $\mathcal{C}^{(2)}$  is treated as the same as an error between  $\mathcal{C}^{(1)}$  and  $\mathcal{C}^{(K)}$ , which is undesirable for ordinal problems.

Furthermore, the loss does not constrain the model to produce unimodal probabilities, so inconsistencies can be produced such as  $\hat{y}_{nj} > \hat{y}_{n\ell} < \hat{y}_{ni}$ , even when  $1 \leq j < \ell < i \leq K$ . Cross-entropy is a fair approach for nominal data, where no additional information is available. By concentrating just on the mode of the distribution and disregarding all other values in the output probability vector, the ordinal information inherent in the data is ignored. However, for ordinal data, the order can be explored to further regularize learning.

**Ordinal Encoding (OE):** Classes are encoded using a cumulative distribution. Therefore,  $y_{nm}$  is represented by 1 if  $k < k_n^*$  and 0 otherwise. Notice that a model should produce class probabilities whose difference should tend to be incremental since it represents a cumulative distribution. This way of introducing ordinality has the advantage of being independent of the way the model is optimized. During inference, to convert back the cumulative distribution to the individual probability of a class  $k$ , the differences between classes should be produced  $p_{nk} = y_{nk} - y_{nk-1}$ , except for the first class [17,18].

**Unimodal (U):** Constraining the classification output to follow a certain discrete ordinal probability (e.g., binomial or Poisson) may be another approach of promoting ordinality.

- **Binomial Unimodal (BU):** This directly constrains the neural network's output, tackling the problem as a regression problem. Instead of several outputs, a single output is predicted, representing the class being predicted. That is,  $y_n = 0$  represents  $k_n^* = 1$  and  $y_n = 1$  represents  $k_n^* = K$  [16]. This model has a single output as the final layer which uses a sigmoid activation function to convert it into class probabilities using the binomial probability mass function. This parametric function which is applied to the output probabilities attempts to maintain the ordinality of the classes.
- **Poisson Unimodal (PU):** Similarly, the constraint for a discrete unimodal probability is enforced by a Poisson probability mass function [14]. On the final layer, the log Poisson probability mass function is applied, followed by a softmax activation function to normalize the output as a probability distribution. According to the authors, it tends to not work as well for more than eight labels [14].

Furthermore, [14] also aims to control the variance of the distribution through a learnable softmax temperature term ( $\tau$ ). In our experiments, a constant value of  $\tau = 1$  was used.

To ensure the ordinality assumption, these parametric techniques may sacrifice accuracy. This tradeoff may be too much in some cases, especially considering the size of the modern deep learning datasets and the high number of mislabeled samples. CE has limitations when applied to ordinal data, as previously stated. By concentrating just on the mode of the distribution and disregarding all other values in the output probability vector, the ordinal information inherent in the data is ignored.

**CO and CO2 Ordinal Losses:** A regularization term that penalizes deviations from the unimodal setting is added to CE [2].

Defining  $\text{ReLU}(x) = \max(0, x)$ , a possible fix for an order-aware loss has been previously proposed as

$$\text{CO2}(\delta, \mathbf{y}_n, \hat{\mathbf{y}}_n) = \text{CE}(\mathbf{y}_n, \hat{\mathbf{y}}_n) + \lambda u(\delta, \mathbf{y}_n, \hat{\mathbf{y}}_n). \quad (3)$$

where  $\lambda \geq 0$  controls the relative influence of the extra term  $u$  which favors unimodal distributions and is defined as

$$u(\delta, \mathbf{y}_n, \hat{\mathbf{y}}_n) = \sum_{k=1}^{k_n^*} \text{ReLU}(\delta + \hat{y}_{nk} - \hat{y}_{n(k+1)}) + \sum_{k=k_n^*}^{K-1} \text{ReLU}(\delta + \hat{y}_{n(k+1)} - \hat{y}_{nk}). \quad (4)$$

Furthermore, a margin of  $\delta > 0$  ensures that the difference between consecutive probabilities is at least  $\delta$  [2]. A value of  $\delta = 0.05$  has been empirically found to provide a sensible margin. As a special case, CO has been defined as the case when the margin is zero ( $\delta = 0$ ),

$$\text{CO}(\mathbf{y}_n, \hat{\mathbf{y}}_n) = \text{CO2}(0, \mathbf{y}_n, \hat{\mathbf{y}}_n). \quad (5)$$

**Ordinal Entropy Loss Function (HO2):** Some of the negative aspects of the CE in the CO and CO2 losses have been previously mentioned: It tries to maximize the probability estimated in the true output class while ignoring the remaining probabilities. CO and CO2 promote unimodality, but do not penalize, or barely penalize, flat distributions. A softer assumption is that the distribution should have low entropy [2].

The HO2 tries to tackle these problems by using an entropy term, instead of the cross-entropy term,

$$\text{HO2}(\delta, \mathbf{y}_n, \hat{\mathbf{y}}_n) = H(\hat{\mathbf{y}}_n) + \lambda u(\delta, \mathbf{y}_n, \hat{\mathbf{y}}_n), \quad (6)$$

where  $H(\mathbf{p})$  denotes the entropy of the distribution  $\mathbf{p}$ .

### 3. Proposal

While the HO2 loss mitigates some negative aspects of using CE in the CO and CO2 loss function, it may be too restrictive to penalize probabilities associated with classes that are further from the true class. The motivation for our proposal (quasi-unimodal loss function (QUL)) stems from the definition of the ordinal setting in [19,20], where it is suggested that in a model consistent with the ordinal setting, a small change in the input data should not lead to a “big jump” in the output decision. Assuming  $f(x)$  as a decision rule that assigns each value of  $x$  to the index  $\in \{1, 2, \dots, K\}$  of the predicted class, the decision rule is said to be consistent with an ordinal data classification setting in a point  $\mathbf{x}_0$  only if  $\exists \varepsilon > 0 \forall \mathbf{x} \in \mathcal{V}_\varepsilon(\mathbf{x}_0) \max f(\mathbf{x}) - \min f(\mathbf{x}) \leq 1$ , with  $\mathcal{V}_\varepsilon$  representing a neighborhood with a small perturbation ( $\varepsilon$ ). Equivalently, the decision boundaries in the input space  $\mathbf{x}$  should be only between regions of consecutive classes. Note that the concept of consistency with the ordinal setting is independent of the type of model (probabilistic or not) and relies only on the decision region output by the model.

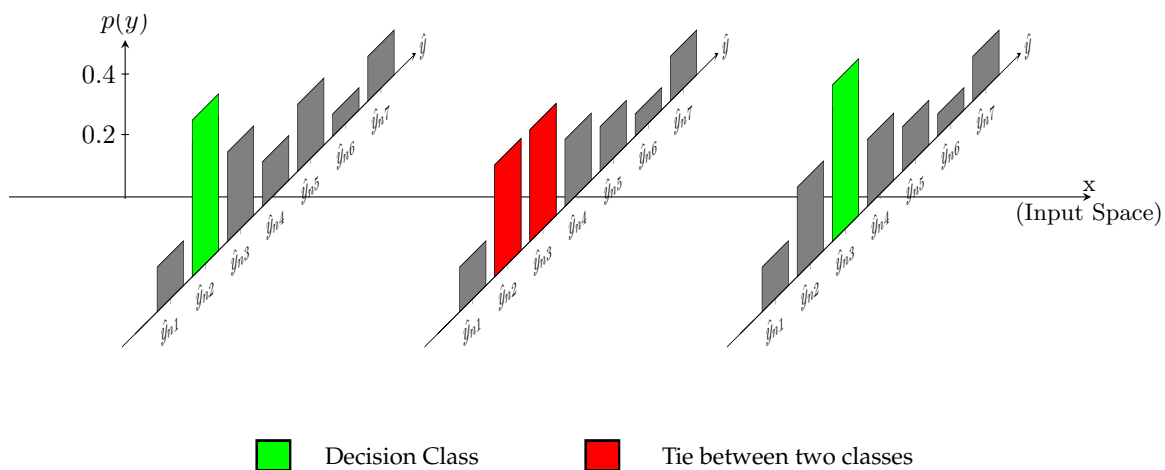
It is easy to conclude that a continuous model  $g(\mathbf{x})$  that outputs a unimodal distribution in the  $K - 1$  dimensional probability simplex is ordinal consistent (as long as there are no ties between more than two values in the unimodal distribution). While this is a sufficient condition, unimodality is not necessary to achieve ordinal consistency (Figure 2).

Let  $y(k)$  be a discrete distribution over the set  $\{1, \dots, K\}$ , with  $k^* = \arg_k \max y(k)$ . We say that  $y(k)$  is a quasi-unimodal distribution (QUD) if:

$$y(k^* + 1) \geq y(k), \quad k \in \{k^* + 2, \dots, K\}, \quad (7)$$

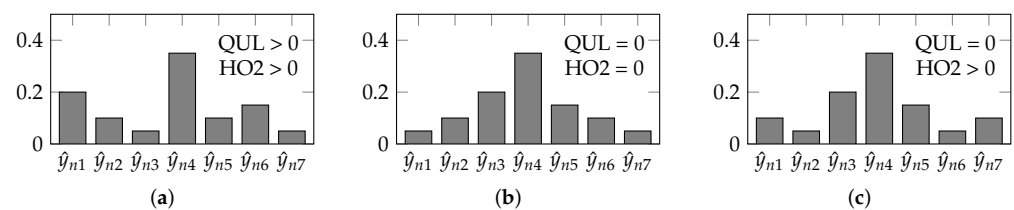
$$y(k) \leq y(k^* - 1), \quad k \in \{1, \dots, k^* - 2\}. \quad (8)$$

It should also be clear that a continuous model  $g(\mathbf{x})$  outputting a QUD  $y(k)$  in the  $K - 1$  dimensional probability simplex is ordinal consistent (as long as there are no ties between more than two values in the quasi-unimodal distribution).



**Figure 2.** Schematic representation of an example where there is ordinal consistency in the decision, without a unimodal distribution in the classes. X axis corresponds to the input space ( $x$ ) of the model, and Y axis corresponds to output probabilities ( $p(\hat{y})$ ) of the model. Along  $x$ , an example of a decision region is represented where the model chooses between two different classes with a boundary region between this decision.

These observations suggest the adoption of a more flexible loss during the training of predictive models, promoting only QUD in the output and not the more restrictive unimodal distribution, as illustrated in Figure 3.



**Figure 3.** Probabilities produced by three different models for an observation  $n$ . HO2 forces the entire distribution curve to have decreasing probabilities. On the other hand, QULHO focuses on the top neighborhood, and is not as concerned about the more distant classes. In this example,  $k_n^* = 4$  is assumed to be the true class. (a) Non-unimodal; (b) Unimodal; (c) Quasi-unimodal.

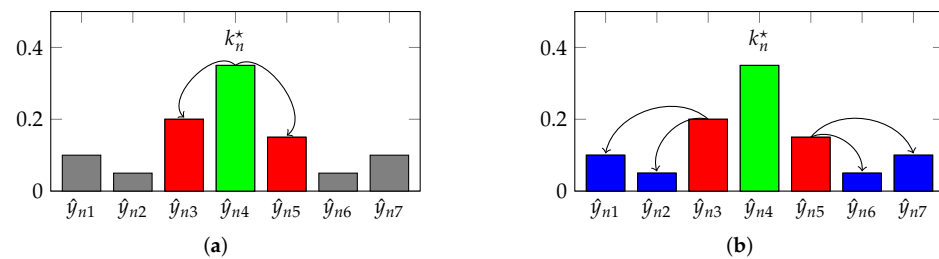
**Quasi-unimodal loss function (QUL):** Compared to HO2, the ordinal terms in the proposal loss promote unimodality only in the top neighborhood output, which allows for a more flexible curve:

$$\begin{aligned} \text{QUL}(\delta, f, \mathbf{y}_n, \hat{\mathbf{y}}_n) = & f(\mathbf{y}_n, \hat{\mathbf{y}}_n) \\ & + \text{ReLU}(\delta + y_{n(k_n^*-1)} - \hat{y}_{nk_n^*}) \\ & + \text{ReLU}(\delta + y_{n(k_n^*+1)} - \hat{y}_{nk_n^*}) \\ & + \lambda \sum_{k=1}^{k_n^*-2} \text{ReLU}(\delta + \hat{y}_{nk} - \hat{y}_{n(k_n^*-1)}) \\ & + \lambda \sum_{k=k_n^*+2}^K \text{ReLU}(\delta + \hat{y}_{nk} - \hat{y}_{n(k_n^*+1)}). \end{aligned} \quad (9)$$

The terms which constrain the top-three neighborhood are in red (Figure 4a), while the terms which constrain the remaining classes relative to the neighborhood are in blue (Figure 4b). Function  $f$  may be defined as CE (like CO2) or as H (like HO2),

$$\text{QULHO}(\delta, \mathbf{y}_n, \hat{\mathbf{y}}_n) = \text{QUL}(\text{H}, \delta, \mathbf{y}_n, \hat{\mathbf{y}}_n). \quad (10)$$

$$\text{QULCE}(\delta, \mathbf{y}_n, \hat{\mathbf{y}}_n) = \text{QUL}(\text{CE}, \delta, \mathbf{y}_n, \hat{\mathbf{y}}_n). \quad (11)$$



**Figure 4.** Schematic representation of the two QUL terms as in Equation (9): The first term penalizes only the direct neighborhood, while the second term penalizes the remaining classes relative to the neighborhood. (a) First QUL term; (b) Second QUL term.

## 4. Methods

### 4.1. Architectures

Convolutional neural networks (CNNs) are used since all datasets consist of images. Filters are learnt in these models, which are quadrilateral patches that are convolved throughout the whole input image—unlike fully-connected networks, just a local neighborhood of inputs is coupled at each layer. Each convolution is usually combined with downsampling procedures, such as max-pooling, which gradually lower the size of the original input image while increasing the receptive field.

Three traditional different CNN architectures are used here: MobileNet\_V2 [21], ResNet18 [22] and VGG-16 [23]. These three architectures are widely used, well known and frequently referenced in the literature. They were pre-trained on ImageNet using PyTorch (<https://pytorch.org/vision/stable/models.html> (accessed on: 10 December 2021)).

The following layers were used to replace the last block of each architecture: Dropout with  $p = 20\%$ , 512-unit dense layer with ReLU, dropout with  $p = 20\%$ , a 256-wide dense layer with ReLU, followed by  $K$  output neurons corresponding to the  $K$  output classes of each dataset.

### 4.2. Losses

Eight different losses are evaluated in this work: Cross Entropy (CE) for the baseline model and seven distinct ordinal losses: Ordinal Encoding (OE), Poisson Unimodal (PU), Binomial Unimodal (BU), CO2, Ordinal Entropy Loss Function (HO2), quasi-unimodal entropy loss function (QULHO) and quasi-unimodal cross-entropy loss function (QULCE).

### 4.3. Training

The weights of the aforementioned architectures were initialized during training using ImageNet pre-training. ADAM was used as the optimizer, with a learning rate of  $10^{-4}$ . When the loss is constant for 10 epochs using a specific scheduler, the learning rate is lowered by 10%. The training process is completed after 100 epochs for the Herlev and AFAD dataset; however, the focupath dataset required 300 epochs to finish the training. The training process was conducted on a Nvidia GTX 1080ti (11GB) GPU and on a Nvidia tesla v100 (32 GB). In the case of the ordinal entropy loss function and quasi-unimodal entropy loss function, the hyperparameter  $\lambda$  is tuned by conducting nested  $k$ -fold cross-validating using the training set (with  $k = 5$ ) to generate an unbiased validation set ( $\lambda \in [0.00001; 0.0001; 0.001; 0.01; 0.1]$ ). A fix value of  $\omega = 0.05$  was used during the training.

As elaborated in the experimental section below, experiments were also performed in which the loss QULHO was initialized by pre-training on HO2 to further facilitate the optimization process.



## 5. Experimental Details

### 5.1. Datasets

Three different datasets are used—Herlev dataset, FocusPath dataset and Asian Face Age Dataset (AFAD)—which will now be elaborated. All these datasets consist of RGB images.

#### 5.1.1. Herlev Dataset

The Herlev dataset collected at the Herlev University Hospital (Denmark) is a publicly accessible dataset (<http://mde-lab.aegean.gr/index.php/downloads> (accessed on: 5 November 2021)) using a digital camera and microscope with an image resolution of  $0.201\ \mu\text{m}$  per pixel [24]. The specimens were prepared using the conventional pap smear and pap staining methods. Two cytotechnicians and a doctor classified the cervical images in the Herlev dataset into seven classes to improve the precision of diagnosis.











































A total of 917 images of individual cervical cells constitute the Herlev dataset. Each image has a classification label and ground truth segmentation. Table 1 illustrates the nomenclature of the seven different classes from the dataset, wherein classes 1–3 correspond to different types of normal cells and classes 4–7 correspond to varying levels of abnormal cells, according to the World Health Organization (WHO) classification system. Furthermore, The Bethesda System (TBS) classification system condenses the seven classes into only four classes, as illustrated in the table.

#### 5.1.2. FocusPath Dataset

The public histopathological dataset entitled FocusPath (<https://zenodo.org/record/3926181#YPFgluhKjIU> (accessed on: 5 November 2021)) contains image quality annotations [25]. The FocusPath dataset contains 8640 patches of  $1024 \times 1024$  images. Nine different stained slides from various human organs were used to create these images. The Huron TissueScope LE12 scanned the original Whole Slide Images. It uses a  $40\times$  optics lens at  $0.25\ \mu\text{m}$ /pixel resolution. For the focus level, there are 14 absolute z-level scores that correspond to the ground-truth class.

Table 2 shows multiple examples from each of the FocusPath dataset's classes. Due to the low quantity of examples of images in the more defocused classes (12 and 13), the label was changed to belong to class 11. This way, the dataset was then divided into 12 different focus classes—0 (focus patch) to 11 (defocus patch).













**Table 2.** Examples of 14 FocusPath classes.

													
													
													
0	1	2	3	4	5	6	7	8	9	10	11	12	13
Focus Level—focus(0) → defocus(13)													

#### 5.1.3. AFAD Dataset

The Asian Face Age Dataset (AFAD) [26] is a publicly available dataset (<https://github.com/afad-dataset/tarball> (accessed on: 10 November 2021)) used to estimate age. Images smaller than  $64 \times 64$  were discarded, resulting in 164,432 photos (63,680 of females and 100,752 of males). The ages range from 15 to 72 for a total of  $K = 57$  classes (Table 3).

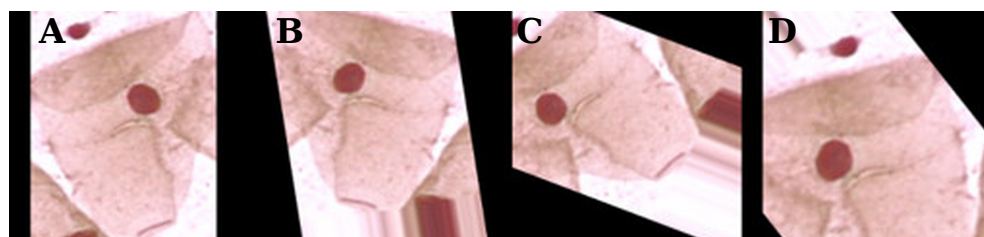
**Table 3.** Examples of six different classes from the AFAD dataset.

					
					
15	25	35	45	55	66
Age (− → +)					

### 5.2. Data Pre-Processing

The three datasets are partitioned into train, validation, and test subsets (60-20-20%), with the ratio between the different classes maintained. To feed the network and to use the pre-trained architectures on ImageNet it was necessary to resize the images to  $224 \times 224$  pixels. As typically done in deep learning models, normalization was also performed to scale the pixel values to 0–1, which speeds up the training process and also helps the network in regularization.

Furthermore, to tackle the small amount of data in the datasets, data augmentation is performed. During training, a series of random transformations are applied to each image: 10% of zoom, 10% of width and height shift, horizontal and vertical flips, and image rotation. These transformations are illustrated in Figure 5.

**Figure 5.** Examples of data augmentation on the Herlev database. The original zero-padding image (A) and random transformations (B–D).

### 5.3. Evaluation Metrics

Four distinct metrics were used to assess the performance of the various models: Accuracy, mean absolute error (MAE), mean squared error (MSE), and the Uniform Ordinal Classification Index (UOC).

Accuracy (ACC) is one of the most commonly used measures in classification tasks. For  $N$  observations, taking  $k_i$  and  $\hat{k}_i$  to be the label and prediction of the  $n$ -th observation, respectively, then  $\text{Acc} = \frac{1}{N} \sum_{n=1}^N \mathbb{1}(\hat{k}_n^* = k_n^*)$ , where  $\mathbb{1}$  is the indicator function.

However, this metric treats all class errors as equal, whether the error is between adjacent classes or between classes in the extreme. If we have  $K$  classes represented by a set  $\mathcal{C} = \{\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \dots, \mathcal{C}^{(K)}\}$ , then accuracy will treat an error between  $\mathcal{C}^{(1)}$  and  $\mathcal{C}^{(2)}$  with the same magnitude as an error between  $\mathcal{C}^{(1)}$  and  $\mathcal{C}^{(K)}$  which is clearly worse. For that reason, a popular metric for ordinal classification is the Mean Absolute Error (MAE),  $\text{MAE} = \frac{1}{N} \sum_i |k_i^* - \hat{k}_i^*|$ . This metric is not perfect since it treats an ordinal variable as a cardinal variable. An error between classes  $\mathcal{C}^{(1)}$  and  $\mathcal{C}^{(3)}$  will be treated as two times worse than an error between classes  $\mathcal{C}^{(1)}$  and  $\mathcal{C}^{(2)}$ . Naturally, the assumption of cardinality is not always warranted. In Mean Squared Error (MSE), instead of being the absolute difference as in MAE, the squared difference is computed between the output of the classifier and the correct label over all examples. This makes the error more pronounced the further away the predicted class is to the true class [27].

Uniform ordinal classification index (UOC) is a recent metric which combines accuracy and ranking in performance assessment and is also robust against imbalanced classes [28].



The better the performance, the lower the UOC. UOC consists of a minimization over the set of all consistent paths that can be traced over the confusion matrix, similar to ordinal classification index (OC) [29]. Each path has a benefit and a penalty, with the benefit being for paths that better follow the natural order of the classes and the penalty being for paths that deviate from the main diagonal, incorporating a classification component into the performance evaluation. Unlike OC, UOC is based on a probabilistic approach, with class priors replaced by a uniform class distribution, making it robust to class imbalance.

By combining these four different evaluation metrics we hope to provide a balanced view of the performance of the methods.

## 6. Results and Discussion

The models' performances for the 10-fold (Herlev and FocusPath datasets) and 1-fold (AFAD dataset) using three distinct architectures and three different datasets are presented in Tables 4–7, with the eight different learning losses—conventional Cross-Entropy (CE), Binomial Unimodal (BU) [16], Poisson Unimodal (PU) [14], Ordinal Encoding (OE) [18], CO2, Ordinal Entropy Loss (HO2) [2], and our proposal losses QULCE and QULHO as measured by MAE, accuracy, MSE and UOC as detailed in the previous section. The best models are shown in **bold**.

**Table 4.** Results for the **Herlev dataset** when considering **seven classes**, averaged for 10 folds.

	Mean Absolute Error (MAE)				Accuracy (%)			
	MobileNet_v2	ResNet18	VGG16	Avg	MobileNet_v2	ResNet18	VGG16	Avg
CE	0.34 ± 0.05	0.34 ± 0.09	0.37 ± 0.09	0.37	75.0 ± 4.4	74.4 ± 6.1	73.1 ± 4.7	73.4
BU	0.36 ± 0.04	0.36 ± 0.06	0.44 ± 0.05	0.41	69.0 ± 3.5	69.5 ± 3.7	63.9 ± 4.6	66.2
PU	<b>0.31 ± 0.04</b>	0.35 ± 0.06	0.44 ± 0.10	0.39	74.2 ± 2.8	73.3 ± 4.3	67.6 ± 6.2	70.1
OE	0.33 ± 0.05	0.35 ± 0.10	0.37 ± 0.06	0.36	74.4 ± 3.8	73.6 ± 6.4	72.6 ± 3.8	72.8
CO2	0.34 ± 0.06	0.34 ± 0.07	0.36 ± 0.06	0.37	73.1 ± 3.7	73.3 ± 4.5	71.8 ± 3.3	72.6
HO2	0.34 ± 0.05	0.35 ± 0.10	0.36 ± 0.07	0.37	74.1 ± 3.9	73.3 ± 6.4	72.0 ± 3.7	72.0
QULCE	0.34 ± 0.06	0.35 ± 0.06	0.36 ± 0.10	0.35	<b>75.2 ± 3.6</b>	73.8 ± 3.1	<b>73.5 ± 4.9</b>	74.2
QULHO	0.34 ± 0.07	<b>0.31 ± 0.07</b>	<b>0.35 ± 0.06</b>	0.33	73.2 ± 5.6	<b>74.5 ± 5.0</b>	73.2 ± 3.9	73.6

	Mean Squared Error (MSE)				UOC (%)			
	MobileNet_v2	ResNet18	VGG16	Avg	MobileNet_v2	ResNet18	VGG16	Avg
CE	0.59 ± 0.13	0.58 ± 0.21	0.64 ± 0.23	0.60	36.0 ± 5.7	36.2 ± 9.3	<b>38.5 ± 8.2</b>	36.9
BU	0.48 ± 0.07	0.47 ± 0.13	0.65 ± 0.12	0.53	39.7 ± 4.9	40.1 ± 5.7	47.2 ± 4.9	42.3
PU	<b>0.44 ± 0.09</b>	0.56 ± 0.14	0.74 ± 0.25	0.58	<b>33.6 ± 4.5</b>	37.2 ± 6.3	45.5 ± 8.6	38.8
OE	0.50 ± 0.10	0.57 ± 0.21	0.61 ± 0.15	0.56	35.4 ± 5.6	37.3 ± 9.1	39.0 ± 6.4	37.2
CO2	0.54 ± 0.15	0.55 ± 0.18	0.57 ± 0.15	0.55	36.2 ± 6.4	37.1 ± 7.6	40.2 ± 6.1	37.9
HO2	0.54 ± 0.12	0.57 ± 0.24	<b>0.56 ± 0.18</b>	0.55	36.2 ± 6.1	37.8 ± 8.7	39.6 ± 6.8	37.9
QULCE	0.58 ± 0.16	0.60 ± 0.17	0.62 ± 0.29	0.60	36.3 ± 5.1	37.1 ± 5.1	38.6 ± 8.4	37.3
QULHO	0.52 ± 0.16	<b>0.45 ± 0.13</b>	0.57 ± 0.14	0.51	36.5 ± 7.3	<b>34.7 ± 6.9</b>	38.8 ± 5.3	36.7

**bold:** best model.

In general, results are fairly close—except that the proposed losses sometimes having a high variance due to them being harder to optimize. For that reason, we introduce a version that is pre-trained from the HO2 loss, in order to mitigate such optimization problems.

Even so, the proposed loss is quite competitive, especially on the Herlev dataset. Please notice that BU and PU fail to run as the number of classes increases after a certain point.

In the results for the AFAD dataset, due to a large number of classes, HO2 and QULHO losses have optimization problems, which suggests that a more exhaustive fine-tuning of hyper-parameters must be done.

The present work intends to promote the use of ordinal losses against nominal losses when the problem at hand has ordinal data. In the Herlev and FocusPath dataset in most of the cases ordinal losses, across the four different metrics, had better results than nominal cross-entropy loss.

**Table 5.** Results for the **Herlev dataset** when considering **four classes**, averaged for 10 folds.

	Mean Absolute Error (MAE)				Accuracy (%)			
	MobileNet_v2	ResNet18	VGG16	Avg	MobileNet_v2	ResNet18	VGG16	Avg
CE	0.22 ± 0.06	0.24 ± 0.03	0.27 ± 0.06	0.25	<b>81.8 ± 4.3</b>	79.8 ± 2.6	77.9 ± 4.8	79.2
BU	<b>0.21 ± 0.03</b>	0.26 ± 0.05	0.28 ± 0.06	0.27	80.7 ± 2.5	77.2 ± 2.3	74.4 ± 4.7	76.5
PU	0.24 ± 0.05	0.24 ± 0.05	0.26 ± 0.05	0.25	78.8 ± 3.4	78.5 ± 4.1	77.5 ± 3.8	77.8
OE	0.22 ± 0.06	<b>0.22 ± 0.04</b>	<b>0.24 ± 0.03</b>	0.24	81.2 ± 4.9	<b>80.7 ± 4.1</b>	<b>79.4 ± 2.5</b>	79.2
CO2	0.24 ± 0.05	0.22 ± 0.04	0.26 ± 0.05	0.25	79.2 ± 3.2	80.4 ± 3.8	77.0 ± 3.9	78.5
HO2	0.22 ± 0.05	0.26 ± 0.06	0.27 ± 0.05	0.26	80.8 ± 3.7	78.0 ± 4.3	77.4 ± 3.7	77.8
QULCE	0.24 ± 0.03	0.25 ± 0.05	0.28 ± 0.07	0.25	79.6 ± 2.8	79.2 ± 3.8	77.9 ± 4.8	78.9
QULHO	0.24 ± 0.06	0.24 ± 0.05	0.25 ± 0.03	0.25	79.2 ± 4.5	79.9 ± 4.3	78.6 ± 1.7	79.2
	Mean Squared Error (MSE)				UOC (%)			
	MobileNet_v2	ResNet18	VGG16	Avg	MobileNet_v2	ResNet18	VGG16	Avg
CE	0.33 ± 0.10	0.31 ± 0.05	0.37 ± 0.12	0.34	30.1 ± 6.9	31.4 ± 4.6	35.3 ± 6.4	32.2
BU	<b>0.26 ± 0.05</b>	0.33 ± 0.11	0.33 ± 0.08	0.31	<b>29.2 ± 3.7</b>	33.1 ± 3.7	36.2 ± 6.4	32.8
PU	0.30 ± 0.08	0.30 ± 0.08	0.32 ± 0.09	0.31	32.8 ± 5.2	32.3 ± 5.5	34.6 ± 4.7	33.2
OE	0.30 ± 0.11	<b>0.28 ± 0.06</b>	<b>0.30 ± 0.06</b>	0.29	30.6 ± 7.5	<b>29.4 ± 6.0</b>	<b>32.3 ± 3.8</b>	30.8
CO2	0.30 ± 0.11	0.28 ± 0.07	0.33 ± 0.08	0.31	32.5 ± 5.5	30.3 ± 4.1	34.7 ± 5.5	32.5
HO2	0.29 ± 0.10	0.34 ± 0.10	0.36 ± 0.11	0.33	30.5 ± 5.4	33.2 ± 6.7	35.1 ± 6.0	32.9
QULCE	0.32 ± 0.05	0.33 ± 0.08	0.41 ± 0.13	0.35	31.9 ± 3.8	31.8 ± 5.2	35.4 ± 7.8	33.0
QULHO	0.31 ± 0.10	0.34 ± 0.06	0.34 ± 0.06	0.33	32.5 ± 6.1	31.8 ± 6.0	33.6 ± 3.1	32.6

**bold:** best model.**Table 6.** Results for the **FocusPath dataset**, averaged for 10 folds.

	Mean Absolute Error (MAE)				Accuracy (%)			
	MobileNet_v2	ResNet18	VGG16	Avg	MobileNet_v2	ResNet18	VGG16	Avg
CE	0.10 ± 0.02	<b>0.11 ± 0.02</b>	0.16 ± 0.02	0.12	91.0 ± 2.0	<b>89.3 ± 1.4</b>	85.4 ± 1.6	88.6
PU	—	—	—	—	—	—	—	—
BU	0.20 ± 0.03	0.18 ± 0.03	0.49 ± 0.96	0.29	80.6 ± 2.4	82.1 ± 3.2	75.6 ± 22.8	79.4
OE	<b>0.09 ± 0.01</b>	0.12 ± 0.02	<b>0.13 ± 0.02</b>	0.11	<b>91.4 ± 1.4</b>	88.6 ± 1.9	<b>87.5 ± 1.2</b>	89.2
CO2	0.11 ± 0.03	0.13 ± 0.02	0.18 ± 0.02	0.14	89.3 ± 2.2	87.3 ± 1.9	83.1 ± 1.5	86.6
HO2	0.13 ± 0.02	0.14 ± 0.04	0.21 ± 0.02	0.16	87.7 ± 2.1	86.8 ± 3.4	81.5 ± 1.6	85.3
QULCE	0.13 ± 0.01	0.14 ± 0.05	0.20 ± 0.02	0.16	87.8 ± 1.7	87.3 ± 3.7	81.5 ± 1.9	85.6
QULHO	0.12 ± 0.02	0.13 ± 0.04	0.19 ± 0.01	0.15	88.7 ± 1.6	87.2 ± 3.9	82.8 ± 1.4	86.3
	Mean Squared Error (MSE)				UOC (%)			
	MobileNet_v2	ResNet18	VGG16	Avg	MobileNet_v2	ResNet18	VGG16	Avg
CE	0.13 ± 0.09	0.15 ± 0.10	0.22 ± 0.10	0.17	17.5 ± 3.3	<b>19.4 ± 3.4</b>	29.1 ± 3.2	22.0
BU	0.20 ± 0.03	0.20 ± 0.06	0.18 ± 0.03	0.20	31.6 ± 3.5	27.1 ± 4.8	29.4 ± 2.5	29.4
PU	4.96 ± 9.70	0.15 ± 0.06	24.38 ± 0.06	9.83	36.6 ± 31.8	21.7 ± 4.8	99.8 ± 0.0	52.7
OE	<b>0.09 ± 0.02</b>	<b>0.13 ± 0.03</b>	<b>0.15 ± 0.05</b>	0.12	<b>15.9 ± 3.6</b>	21.4 ± 3.0	<b>27.0 ± 4.1</b>	21.4
CO2	0.13 ± 0.05	0.15 ± 0.04	0.22 ± 0.06	0.17	19.9 ± 4.7	25.0 ± 5.0	33.9 ± 6.4	26.2
HO2	0.15 ± 0.03	0.19 ± 0.14	0.28 ± 0.10	0.21	27.8 ± 6.0	26.0 ± 6.8	37.5 ± 6.3	30.4
QULCE	4.38 ± 1.66	1.35 ± 2.52	0.53 ± 0.32	2.09	84.1 ± 24.2	35.0 ± 29.0	44.7 ± 14.4	54.6
QULHO	0.14 ± 0.06	0.17 ± 0.11	0.24 ± 0.05	0.18	23.9 ± 4.9	25.2 ± 7.3	34.4 ± 4.9	27.8

**bold:** best model.

**Table 7.** Results for the AFAD dataset, averaged for three folds.

	Mean Absolute Error (MAE)				Accuracy (%)			
	MobileNet_v2	ResNet18	VGG16	Avg	MobileNet_v2	ResNet18	VGG16	Avg
CE	0.02 ± 0.01	<b>0.01 ± 0.00</b>	<b>0.02 ± 0.00</b>	0.02	99.2 ± 0.4	<b>99.7 ± 0.0</b>	<b>99.4 ± 0.0</b>	99.5
PU	—	—	—	—	—	—	—	—
BU	—	—	—	—	—	—	—	—
OE	<b>0.01 ± 0.01</b>	0.04 ± 0.03	0.03 ± 0.01	0.03	<b>99.3 ± 0.5</b>	96.8 ± 2.1	97.8 ± 0.3	98.0
CO2	0.42 ± 0.31	0.24 ± 0.24	0.16 ± 0.15	0.27	84.7 ± 7.4	95.6 ± 2.0	95.8 ± 2.1	92.0
HO2	1.84 ± 1.28	2.99 ± 0.01	3.01 ± 0.65	2.61	66.0 ± 16.1	55.3 ± 0.0	53.9 ± 6.6	58.4
QULCE	0.17 ± 0.02	0.03 ± 0.00	0.02 ± 0.00	0.07	97.0 ± 0.3	99.1 ± 0.0	99.3 ± 0.0	98.5
QULHO	1.62 ± 1.34	2.98 ± 0.01	2.94 ± 0.67	2.51	76.3 ± 19.9	55.4 ± 0.0	55.1 ± 6.0	62.3
	Mean Squared Error (MSE)				UOC (%)			
	MobileNet_v2	ResNet18	VGG16	Avg	MobileNet_v2	ResNet18	VGG16	Avg
CE	0.14 ± 0.08	<b>0.03 ± 0.00</b>	<b>0.09 ± 0.01</b>	0.09	<b>3.7 ± 2.6</b>	<b>0.4 ± 0.0</b>	<b>1.2 ± 0.1</b>	1.8
PU	—	—	—	—	—	—	—	—
BU	—	—	—	—	—	—	—	—
OE	<b>0.04 ± 0.01</b>	0.11 ± 0.06	0.11 ± 0.02	0.09	8.4 ± 5.2	41.8 ± 13.4	7.8 ± 2.9	19.3
CO2	3.80 ± 3.35	3.01 ± 3.08	1.81 ± 2.02	2.87	81.2 ± 22.7	93.3 ± 2.3	64.6 ± 38.3	79.7
HO2	19.10 ± 13.40	30.40 ± 0.09	30.49 ± 9.47	26.66	98.2 ± 0.9	99.3 ± 0.0	99.3 ± 0.1	98.9
QULCE	1.42 ± 0.22	0.27 ± 0.02	0.18 ± 0.01	0.62	12.0 ± 3.9	8.9 ± 8.4	10.0 ± 8.3	10.3
QULHO	17.65 ± 13.81	30.34 ± 0.07	29.86 ± 9.72	25.95	97.8 ± 1.2	99.3 ± 0.0	99.2 ± 0.1	98.8

**bold:** best model.

## 7. Conclusions

In ordinal classification, metrics are most concerned about the relative order of the predictions, rather than the absolute order. That is, an error between two distant classes is deemed worse than an error between two neighboring classes. For that reason, a wide variety of losses can be found in the literature.

In this work, a novel non-parametric ordinal loss is proposed, which induces output probabilities to follow a quasi-unimodal distribution. Other strategies in the literature involve large modifications in architecture or data format; therefore, the suggested loss is a handy way to introduce ordinality in the classification task without requiring substantial changes in design or data format. It improves upon previous work by allowing for a more flexible distribution since the unimodal constraint is relaxed for other classes outside the top-three neighborhood. Furthermore, a review of existing ordinal losses is provided.

Several losses from the literature, plus the proposed one, are compared in an empirical assessment which comprises three different datasets using three popular neural network architectures. The parametric losses that were evaluated fail after the number of classes are increased beyond a certain point. The proposed loss seems competitive in terms of scalability and consistency.

**Author Contributions:** Conceptualization, T.A., R.C. and J.S.C.; methodology, T.A., R.C. and J.S.C.; validation, T.A., R.C.; formal analysis, T.A., R.C. and J.S.C.; investigation, T.A., R.C. and J.S.C.; writing—original draft preparation, T.A., R.C. and J.S.C.; writing—review and editing, T.A., R.C. and J.S.C.; visualization, T.A., R.C. and J.S.C.; supervision, J.S.C.; project administration, J.S.C.; funding acquisition, J.S.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** The project TAMI-Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) partially funding this work is co-financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program-NORTE 2020 and by the Portuguese Foundation for Science and Technology-FCT under the CMU-Portugal International Partnership. Tomé Albuquerque was supported by Ph.D. grant 2021.05102.BD, also provided by FCT.

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All the datasets used in this work are publicly available: Herlev dataset—<http://mde-lab.aegean.gr/index.php/downloads> (accessed on: 5 November 2021); Focus-Path dataset—<https://zenodo.org/record/3926181#.YPFgluhKjIU> (accessed on: 5 November 2021); AFAD dataset—<https://github.com/afad-dataset/tarball> (accessed on: 10 November 2021)

**Acknowledgments:** The authors would like to acknowledge access to the Herlev Pap smear dataset collected by Herlev University Hospital (Denmark) and the Technical University of Denmark.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

ACC	Accuracy
BU	Binomial Unimodal
CE	Cross-entropy
CNN	Convolutional Neural Network
MAE	Mean Average Precision
MSE	Mean Squared Error
OE	Ordinal Encoding
PU	Poisson Unimodal
QUL	Quasi-unimodal loss function
QULHO	Quasi-unimodal entropy loss function
QULCE	Quasi-unimodal cross-entropy loss function
TBS	The Bethesda System
WHO	World Health Organization

## References

1. Belharbi, S.; Ayed, I.B.; McCaffrey, L.; Granger, E. Non-parametric Uni-modality Constraints for Deep Ordinal Classification. *arXiv* **2020**, arXiv:1911.10720.
2. Albuquerque, T.; Cruz, R.; Cardoso, J.S. Ordinal losses for classification of cervical cancer risk. *PeerJ Comput. Sci.* **2021**, *7*, e457. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Liu, H.; Lu, J.; Feng, J.; Zhou, J. Ordinal Deep Feature Learning for Facial Age Estimation. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 157–164. [\[CrossRef\]](#)
4. Pan, H.; Han, H.; Shan, S.; Chen, X. Mean-Variance Loss for Deep Age Estimation from a Face. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5285–5294. [\[CrossRef\]](#)
5. Zhu, H.; Zhang, Y.; Li, G.; Zhang, J.; Shan, H. Ordinal Distribution Regression for Gait-based Age Estimation. *Sci. China Inf. Sci.* **2020**, *63*, 120102. [\[CrossRef\]](#)
6. Crammer, K.; Singer, Y. Pranking with Ranking. In *Advances in Neural Information Processing Systems*; Dietterich, T., Becker, S., Ghahramani, Z., Eds.; MIT Press: Cambridge, MA, USA, 2002; Volume 14.
7. Koren, Y.; Sill, J. OrdRec: An ordinal model for predicting personalized item rating distributions. In Proceedings of the Fifth ACM Conference on Recommender Systems, Chicago, IL, USA, 23–27 October 2011; pp. 117–124. [\[CrossRef\]](#)
8. Gentry, A.; Jackson-Cook, C.; Lyon, D.; Archer, K. Penalized Ordinal Regression Methods for Predicting Stage of Cancer in High-Dimensional Covariate Spaces. *Cancer Inform.* **2015**, *14*, 201–208. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Moody, J.E.; Utans, J. Architecture Selection Strategies for Neural Networks: Application to Corporate Bond Rating Predicti. In Proceedings of the Neural Networks in the Capital Markets, NIPS 1995, Denver, CO, USA, 27 November–2 December 1995.
10. Jia, X.; Zheng, X.; Li, W.; Zhang, C.; Li, Z. Facial Emotion Distribution Learning by Exploiting Low-Rank Label Correlations Locally. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9833–9842. [\[CrossRef\]](#)
11. Xiong, H.; Liu, H.; Zhong, B.; Fu, Y. Structured and Sparse Annotations for Image Emotion Distribution Learning. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 363–370. [\[CrossRef\]](#)
12. Zhou, Y.; Xue, H.; Geng, X. Emotion Distribution Recognition from Facial Expressions. In Proceedings of the MM '15: Proceedings of the 23rd ACM International Conference on Multimedia, New York, NY, USA, 26–30 October 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 1247–1250. [\[CrossRef\]](#)
13. Palermo, F.; Hays, J.; Efros, A. Dating Historical Color Images. In *European Conference on Computer Vision*; Springer: Heidelberg, Germany, 2012; Volume 7577, pp. 499–512. [\[CrossRef\]](#)

14. Beckham, C.; Pal, C. Unimodal Probability Distributions for Deep Ordinal Classification. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Precup, D., Teh, Y.W., Eds.; PMLR: Sydney, Australia, 2017; Volume 70, pp. 411–419.
15. Cardoso, J.S.; Pinto da Costa, J.F. Learning to Classify Ordinal Data: The Data Replication Method. *J. Mach. Learn. Res.* **2007**, *8*, 1393–1429.
16. Costa, J.; Cardoso, J. Classification of Ordinal Data Using Neural Networks. In *European Conference on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 690–697. [\[CrossRef\]](#)
17. Frank, E.; Hall, M. A simple approach to ordinal classification. In *European Conference on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 145–156. [\[CrossRef\]](#)
18. Cheng, J.; Wang, Z.; Pollastri, G. A neural network approach to ordinal regression. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1279–1284.
19. Cardoso, J.S.; Sousa, R. Classification Models with Global Constraints for Ordinal Data. In Proceedings of the 2010 Ninth International Conference on Machine Learning and Applications, Washington, DC, USA, 12–14 December 2010; pp. 71–77. [\[CrossRef\]](#)
20. Sousa, R.; Cardoso, J.S. Ensemble of decision trees with global constraints for ordinal classification. In Proceedings of the 2011 11th International Conference on Intelligent Systems Design and Applications, Córdoba, Spain, 22–24 November 2011; pp. 1164–1169. [\[CrossRef\]](#)
21. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
23. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
24. Jantzen, J.; Dounias, G. Analysis of Pap-smear image data. In Proceedings of the Nature-Inspired Smart Information Systems 2nd Annual Symposium, Austria, 29 November–1 December 2006; Volume 10.
25. Hosseini, M.S.; Zhang, Y.; Plataniotis, K.N. Encoding Visual Sensitivity by MaxPol Convolution Filters for Image Sharpness Assessment. *IEEE Trans. Image Process.* **2019**, *28*, 4510–4525. [\[CrossRef\]](#)
26. Niu, Z.; Zhou, M.; Wang, L.; Gao, X.; Hua, G. Ordinal Regression With Multiple Output CNN for Age Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
27. Gaudette, L.; Japkowicz, N. Evaluation Methods for Ordinal Classification. In *Advances in Artificial Intelligence*; Gao, Y., Japkowicz, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 207–210.
28. Silva, W.; Pinto, J.R.; Cardoso, J.S. A Uniform Performance Index for Ordinal Classification with Imbalanced Classes. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8.
29. Cardoso, J.; Sousa, R. Measuring the Performance of Ordinal Classification. *Int. J. Pattern Recognit. Artif. Intell.* **2011**, *25*, 1173–1195. [\[CrossRef\]](#)