Ordinal Regression with Neuron Stick-breaking for Medical Diagnosis

Xiaofeng Liu[1,2][0000−0002−4514−2016]∗, Yang Zou[1][0000−0003−0396−7850]∗, Yuhang Song[3][0000−0003−4990−2964], Chao Yang[3][0000−0002−6553−7963], and Jane You[4][0000−0002−8181−4836] B.V.K Vijaya Kumar[1,5][0000−0001−7126−6381]

[1] Carnegie Mellon University PA 15213, USA
liuxiaofeng@cmu.edu, yzou2@andrew.cmu.edu
[2] Fanhan Information Tech, Suzhou, China
[3] University of Southern California CA 90089, USA
yuhangso@usc.edu,chaoy@usc.edu
[4] The Hong Kong Polytechnic University
csyjia@comp.polyu.edu.hk
[5] Carnegie Mellon University Africa, Kigali, Rwanda
kumar@ece.cmu.edu
∗These two authors contribute equally.

Abstract. The classification for medical diagnosis usually involves inherently ordered labels corresponding to the level of health risk. Previous multi-task classifiers on ordinal data often use several binary classification branches to compute a series of cumulative probabilities. However, these cumulative probabilities are not guaranteed to be monotonically decreasing. It also introduces a large number of hyper-parameters to be fine-tuned manually. This paper aims to eliminate or at least largely reduce the effects of those problems. We propose a simple yet efficient way to rephrase the output layer of the conventional deep neural network. We show that our methods lead to the state-of-the-art accuracy on Diabetic Retinopathy dataset and Ultrasound Breast dataset with very little additional cost.

Keywords: Medical Diagnosis · Ordinal Regression · Deep Neural Network · Stick-breaking · Unimodal Label Smoothing

1 Introduction

Recent advances in deep neural networks (DNN) for natural image tasks have prompted a surge of interest in adapting similar models to medical images [1,2,3]. However, some of the special characteristics of medical diagnosis have, in our opinion, not been anfficiently explored.

The classes of a medical image usually represent the health risk levels, which are inherently ordered. For instance, the Diabetic Retinopathy Diagnosis (DR) involves five levels: no DR (1), mild DR (2), moderate DR (3), severe DR (4) and proliferative DR (5) [4,5]. The Breast Imaging-Reporting and Data System

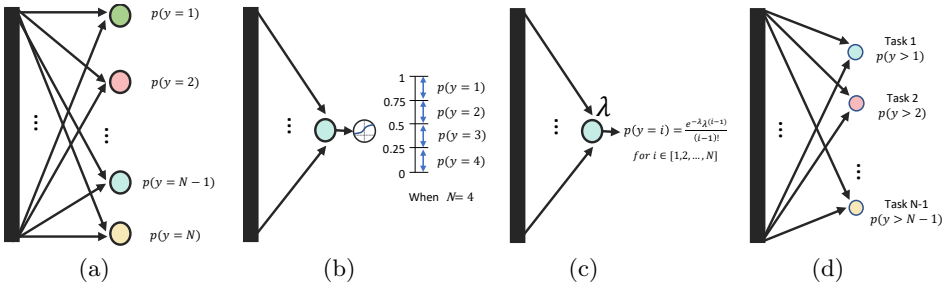

Fig. 1. The architecture of output layer used in previous ordinal regression methods: (a)multi-class classification, (b)regression, (c)Poisson, and (d)multi-task classification. We learn a discriminative mapping from sample $\mathbf{x}$ to an ordinal variable y.

(BIRADS) also includes five diagnostic labels: 1-healthy, 2-benign, 3-probably benign, 4-may contain malignant and 5-probably contains malignant [1,6]. Similar ordinal labeling systems for liver (LIRADS), gynecology (GIRADS), colonography (CRADS) have been established soon afterward [2].

Surely, the ordinal data is not unique to the medical image classification. Some other examples of ordinal labels include the age of a person [7], face expression intensity [8], aesthetic [9], star rating of a movie [10], etc., and are traditionally referred to ordinal regression tasks [11]. Two of the most straightforward approaches either cast it as a multi-class classification problem [12] and optimize the cross-entropy (CE) loss or treat it as a metric regression problem [13] and minimize the absolute/squared error loss (i.e.,MAE/MSE). The former (Fig. 1(a)) assumes that the classes are independent of each other, which totally fails to explore the inherent ordering between the labels. The latter (Fig. 1(c)) treats the discrete labels as continuous numerical values, in which the adjacent classes are equally distant. This assumption violates the non-stationary property of many image related tasks, easily resulting in over-fitting [14].

Recently, better results were achieved via a $N-1$ binary classification subtasks (Fig. 1(b)) using sigmoid output with MSE loss [11] or softmax output with CE loss [2,6,15,16], when we have N levels as the class label. We can transform N levels to a series of labels of length $N-1$. Then the first class is [0,...,0], followed by the second class [1,...,0], third class [1,1,..,0] and so forth. The sub-branches in Fig. 1(b) calculate the cumulative probability $p(y > i|\mathbf{x})$, where i index the class[6]. With the cumulative probability, then it is trivial to define the corresponding discrete probabilities $p(y = i|\mathbf{x})$ via subtraction. These techniques are closely related to their non-deep counterparts [17,18]. However, the cumulative probabilities $p(y > 1|\mathbf{x}), ..., p(y > N-1|\mathbf{x})$ are calculated by several branches independently, therefore, can not guarantee they are monotonically decreasing. That leads to the $p(y = i|\mathbf{x})$ are not guaranteed to be strictly positive and results poor learning efficiency in the early stage of training. Moreover, $N-1$ weights need to be manually fine-tuned to balance the CE loss of each branch.

[6] We will always index probabilities from zero for the remainder of this paper.

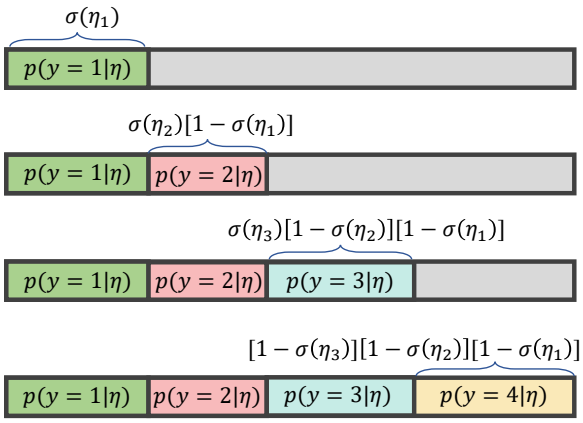

Fig. 2. The Stick-breaking process for 4 classes with 3 boundaries. In [24], η is the linear projection in LGMs.

Besides, under the one-hot target label encoding, the CE loss $\texttt{-log}(p(y = l|\mathbf{x}))$ essentially only cares about the ground truth class l. [19] argues that misclassifying an adult as a baby is more severe than misclassifying as a teenager, even if the probabilities of the adult class are the same. [5,20,21] propose to use a single output neuron to calculate a parameter of a unimodal distribution, and strictly require that the $p(y = i|\mathbf{x})$ follows a Poisson or Binomial distribution, but suffers from lacking the ability to control the variance [21]. Since the peak (also the mean and variance) of a Poisson distribution is equal to a designated λ, we can not assign the peak to the first or last class, and its variance is very high when we need the peak in the very later classes.

Furthermore, the agreement rate of the radiologists for a malignancy is usually less than 80%, which results in a noisy labeled dataset [22,23]. Despite the distinction between adjacent labels is often unclear, it is more likely that a well-trained annotator will mislabel a Severe DR (4) sample to Moderate DR (3) rather than No DR (1).

In this paper, we propose to address the issues discussed above. Briefly, we rephrase the conventional softmax-based output layer to the neuron stick-breaking formulations to guarantee the cumulative probabilities are monotonically decreasing. We evaluated our approaches in the context of medical diagnosis on two datasets, and obtained promising results. We note that although the methods shown here were originally developed for medical images, they are essentially applicable to other ordinal regression problems.

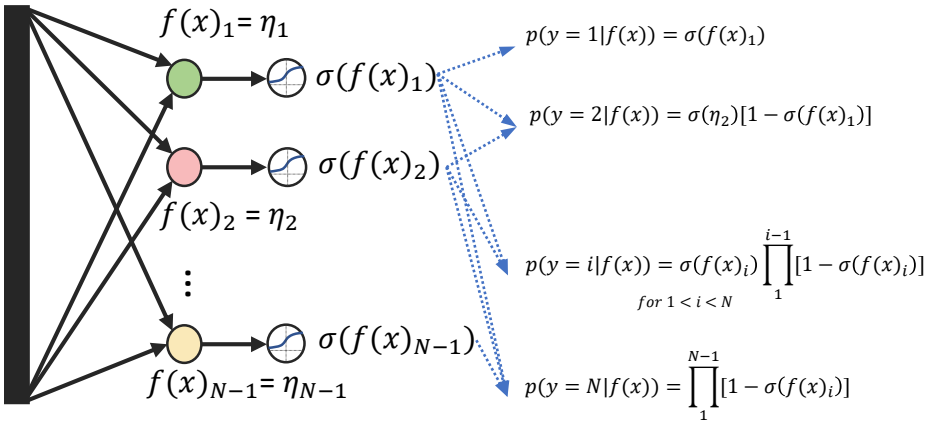

Fig. 3. Our neuron Stick-breaking architecture for N classes with $N-1$ output neurons, followed by sigmoid units and linear operations.

2 Neuron Stick-breaking for ordinal regression

In the stick-breaking approach, we define a stick of unit length between [0,1], and sequentially break off parts of the stick which then become the discrete probabilities for that class (Fig 2(a)) [25]. The stick-breaking process is a subset of the random allocation processes [26] and a generalization of continuation ratio models [27]. It is closely associated with the associated Bayesian non-parametric methods, e.g., [25] used it in constructive definitions of the Dirichlet process [28]. [24] further proposed its parameterization for Latent Gaussian Models (LGMs).

To introduce the stick-breaking processes in a way that is appropriate a deep neural network for ordinal regression, we set $N-1$ output neurons for N levels, and suppose that $f(x)_i$ is a scalar denoting the i-th output of our neural network to substitute linear projections η_i in LGMs. We define the stick length of the first class, i.e., its probability, to be $\sigma(f(x)_1)$, where $\sigma(\cdot)$ denotes the sigmoid nonlinearity. We can then define the second class probability as what was left over from that stick multiplied by the output of the second class, i.e., $(1-\sigma(f(x)_1))\sigma(f(x)_2)$. For the third class probability we compute $(1-\sigma(f(x)_1))(1-\sigma(f(x)_2))\sigma(f(x)_3)$ and so forth, where the last class probability for $p(N|\mathbf{x})$ receives what is left over, i.e., $(1-\sigma(f(x)_1))...(1-\sigma(f(x)_{N-1}))$. The conventional CE loss can be used to train our network[7].

It can be derived that each output $f(x)_i$ is actually the log-ratio $f(x)_i = \texttt{log}(p(y=i|x)/p(y>i|x))$ [24], so these $f(x)_i$ can be interpreted as defining decision boundaries that try to separate the i-th class from all the classes that come after it. By doing so, the prediction is still a discrete probability (i.e.,

[7] Code available at:Anonymous Submission

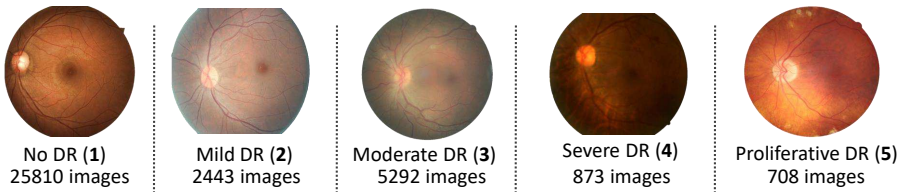

Fig. 4. Some samples with different retinopathy level in the DR dataset.

$\sum_i^{N=1} p(y=i)=1$), and each $p(y=i) \geq 0$, then we do guarantee the relationship of $p(y>1) \geq p(y>2) \geq p(y>N-1)$.

A nice property of our method is that unlike the approaches that only output a single distribution parameter [5,21,29], we obtain a slightly more expressive model since each boundary of two adjacent classes gets its own scalar output $f(x)_i$. The discrete probabilities can also be calculated via our predefined linear manipulations instead of having to estimate cumulative probabilities first [11,17,18]. Therefore, the weights of each branch in [11] are no longer necessary.

3 Experiments

3.1 Datasets

We make use of two typical ordinal datasets in the medical area suitable for DNN implementations. The first dataset contains images of Diabetic Retinopathy (DR) [8]. In this dataset, a large amount of high-resolution fundus (i.e., interior surface at the back of the eye) images data have been labeled as five levels of DR, with levels 1 to 5 representing the No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR, respectively. The left and right fundus image from 17563 patients are publicly available. Following the setting in [21], we adopt the subject-independent ten-fold cross-validation, i.e., the validation set consisting of 10% of the patients is set aside. The images belonging to a patient will only appear in a single fold, in this way we can avoid contamination. The images are also preprocessed as in [5,21] and subsequently resized as 256×256 size images. Some examples can be found in Fig. 4.

The second dataset is the Ultrasound BIRADS (US-BIRADS) [6]. It is comprised of 4904 breast images which are labeled with the BIRADS system. Considering the relatively limited number of samples in level 5, we usually regard the 4-5 as a single level [6]. That results 2700 healthy (1) images, 1113 benign (2) images, 359 probably benign (3), and 732 may contain/contain malignant images. We divide this dataset into 5 subsets for subject-independent five-fold cross validation. We show some samples at different levels in Fig. 5.

[8] https://www.kaggle.com/c/diabetic-retinopathy-detection

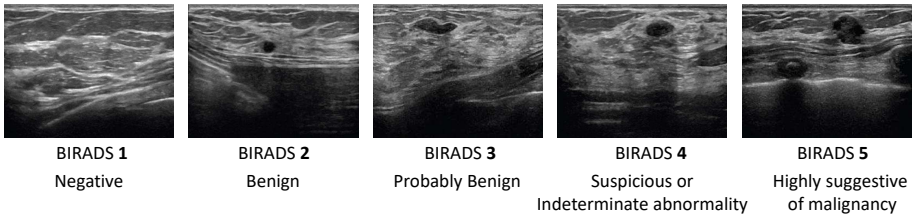

Fig. 5. Some samples with different malignant risk in the US-BIRADS.

3.2 Evaluations

There are several possible evaluation metrics for ordinal data. As a classification problem, the performance of a system can be simply measured by the average classification accuracy. [6] further utilized the Mean True Negative Rate (TNR) at True Positive Rate (TPR) of 0.95. The relatively high TPR used in here is fitted for strict TPR requirement of medical applications to avoid misdiagnosing diseased case as healthy. However, they do not consider the severity of different misclassification. Following the previous metrics in the Kaggle competition of DR dataset, we choose the quadratic weighted kappa (QWK)[9] to implicitly punish the misclassification proportional to the distance between the ground-of-truth label and predicted label of the network [30]. The QWK is formulated as:

$$k = 1 - \frac{\sum_{i,j} \mathbf{W}_{i,j} \mathbf{O}_{i,j}}{\sum_{i,j} \mathbf{W}_{i,j} \mathbf{E}_{i,j}} \tag{1}$$

to measures the level of disagreement between two raters ($\mathcal{A}$ and $\mathcal{B}$). In here, the $\mathcal{A}$ is the $argmax$ prediction of our classifier and $\mathcal{B}$ is the ground truth. The $\mathbf{W}$ is a $N \times N$ matrix where $\mathbf{W}_{i,j}$ denotes the cost associated with misclassifying label i as label j. In QWK, $\mathbf{W}_{i,j} = (i-j)^2$. $\mathbf{O}_{i,j}$ counts the number of images that received a rating i by $\mathcal{A}$ and a rating j by $\mathcal{B}$. The quadratic calculation is one possible choice and one can plug in other distance metrics into kappa calculation. The matrix of expected ratings $\mathbf{E}$, is calculated, assuming that there is no correlation between rating scores. As a result, k is a scalar in [-1,1], and $k = 1$ indicates the two raters are total agreement, whereas $k < 0$ means the classifier performs worse than random choice.

The Mean Absolute Error (MAE) metric is also popular in related ordinal datasets [11], which is computed using the average of the absolute errors between the ground truth and the estimated result. Here, we also propose its use in evaluating the proposed method on two medical ordinal benchmarks.

3.3 Networks and training details

For fair comparison, we choose similar backbones neural networks as in previous works on DP and US-BIRADS datasets. We adjust the last layer and softmax

[9] https://www.kaggle.com/c/diabetic-retinopathy-detection#evaluation

Table 1. Performance on the DR dataset.

Evaluations	Mean TNR@TPR=0.95			Valid Acc	Valid QWK	MAE
	1 *vs* 2-4	1-2 *vs* 3-4	1-3 *vs* 4			
MC	41.5%	30.9%	31.1%	82.4%	0.724	0.37
RG	40.3%	30.6%	30.8 %	76.2%	0.705	0.38
Poisson [21]	38.8%	30.0%	29.6 %	77.1%	0.713	0.38
MT [6]	42.7%	31.7%	31.3%	82.8%	0.726	0.36
NSB	44.0%	33.1%	32.6%	84.2%	0.743	0.32

Table 2. Performance on the US-BIRADS dataset.*Our implementations have slightly higher TNR using MC baseline than the results reported in [6]

.

Evaluations	Mean TNR@TPR=0.95			Valid Acc	Valid QWK	MAE
	1 *vs* 2-5	1-2 *vs* 3-5	1-3 *vs* 4-5			
MC	33.2%*	28.7%*	29.8%*	73.3%	0.678	0.42
RG	31.6%	28.5%	29.5%	73.0%	0.677	0.44
Poisson [21]	29.6%	27.2%	29.5%	72.2%	0.665	0.45
MT [6]	38.5%	29.2%	31.3%	76.5%	0.685	0.41
NSB	39.1%	30.2%	32.0%	78.3%	0.694	0.39

normalization to our neuron stick-breaking formulation. The ResNet [31] style model with 11 ResBlocks as in [21] has been adopted for DR dataset. We use four stick-breaking neurons as our output structure and calculate the $p(y = i|\mathbf{x})$ via the predefined linear operations. AlexNet style architecture [32] with six convolution layers and following two dense layers is used for US-BIRADS image dataset as in [6]. 3 stick-breaking neurons are employed as the last layer. All of networks in our training use the $\mathcal{L}_2$ norm of 10^{-4}, ADAM optimizer [33] with 128 training batch-size and initial learning rate of 10^{-3}. The learning rate will be divided by ten when either the validation loss or the valid set QWK plateaus. We set our hyper-parameters $\eta = 0.15$, $\tau = 1$.

3.4 Numerical Experiments

We conduct our experiments on both datasets with the evaluation metrics discussed earlier. The results in DR dataset are shown in Table 1. Several baseline methods are chosen for comparison, e.g., multi-class classification with CE loss (MC), regression with MSE loss (RG), Poisson distribution output with CE loss (Poisson), and multi-task network with a series of CE loss (MT). The RG is usually worse than MC, but appear to be competitive w.r.t. MAE, since RG optimizes similar metric MSE in its training stage. The Poisson gets the lowest results in the most of evaluations due to its uncontrollable variance. The and MT are more promising than MC as they consider ordinal information. By addressing their limitations, we achieve the state-of-the-art performance in all of the evalu-

ation tasks using the neuron stick-breaking (NSB). The leading performance of our method is also observed on the US-BIRADS dataset (Table 2).

4 Conclusions

We have introduced the stick-breaking presses for DNN-based ordinal regression problem. By reformulating the neurons of the last layer and softmax function, we not only fully consider the ordinal property of the class labels, but also guarantee the cumulative probabilities are monotonically decreasing. We also show how these approaches offer improved performance in DR and US BIRADS datasets. In future work, we intend to leverage our methods for more general ordinal regression tasks.

5 Acknowledgement

This work was supported in part by the National Natural Science Foundation 61308099, 61304032 and 61675202, Hong Kong Government General Research Fund GRF 152202/14E, PolyU Central Research Grant G-YBJW, Youth Innovation Promotion Association, CAS (2017264), Innovative Foundation of CIOMP, CAS (Y586320150), 11ZDGG001,CXJJ-16S038,CXJJ-17S017.

References

1. Geras, K.J., Wolfson, S., Shen, Y., Kim, S., Moy, L., Cho, K.: High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint arXiv:1703.07047 (2017)
2. Li, X., Kao, Y., Shen, W., Li, X., Xie, G.: Lung nodule malignancy prediction using multi-task convolutional neural network. In: Medical Imaging 2017: Computer-Aided Diagnosis. Volume 10134., International Society for Optics and Photonics (2017) 1013424
3. Gentry, A.E., Jackson-Cook, C.K., Lyon, D.E., Archer, K.J.: Penalized ordinal regression methods for predicting stage of cancer in high-dimensional covariate spaces. Cancer informatics **14** (2015) CIN–S17277
4. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama **316**(22) (2016) 2402–2410
5. Beckham, C., Pal, C.: A simple squared-error reformulation for ordinal classification. arXiv preprint arXiv:1612.00775 (2016)
6. Ratner, V., Shoshan, Y., Kachman, T.: Learning multiple non-mutually-exclusive tasks for improved classification of inherently ordered labels. arXiv preprint arXiv:1805.11837 (2018)
7. Eidinger, E., Enbar, R., Hassner, T.: Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security **9**(12) (2014) 2170–2179
8. Zhao, R., Gan, Q., Wang, S., Ji, Q.: Facial expression intensity estimation using ordinal information. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3466–3474

9. Cardoso, J.S., da Costa, J.F.P., Cardoso, M.J.: Modelling ordinal relations with svms: An application to objective aesthetic evaluation of breast cancer conservative treatment. Neural Networks **18**(5-6) (2005) 808–817
10. Koren, Y., Sill, J.: Ordrec: an ordinal model for predicting personalized item rating distributions. In: Proceedings of the fifth ACM conference on Recommender systems, ACM (2011) 117–124
11. 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. (2016) 4920–4928
12. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. IEEE Transactions on pattern analysis and machine intelligence **29**(12) (2007) 2234–2240
13. Fu, Y., Huang, T.S.: Human age estimation with regression on discriminative aging manifold. IEEE Transactions on Multimedia **10**(4) (2008) 578–584
14. Chang, K.Y., Chen, C.S., Hung, Y.P.: Ordinal hyperplanes ranker with cost sensitivities for age estimation. In: Computer vision and pattern recognition (cvpr), 2011 ieee conference on, IEEE (2011) 585–592
15. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2002–2011
16. Chen, S., Zhang, C., Dong, M., Le, J., Rao, M.: Using ranking-cnn for age estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
17. Cheng, J., Wang, Z., Pollastri, G.: A neural network approach to ordinal regression. In: Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, IEEE (2008) 1279–1284
18. Frank, E., Hall, M.: A simple approach to ordinal classification. In: European Conference on Machine Learning, Springer (2001) 145–156
19. Hou, L., Yu, C.P., Samaras, D.: Squared earth mover's distance-based loss for training deep neural networks. arXiv preprint arXiv:1611.05916 (2016)
20. da Costa, J.F.P., Alonso, H., Cardoso, J.S.: The unimodal model for the classification of ordinal data. Neural Networks **21**(1) (2008) 78–91
21. Beckham, C., Pal, C.: Unimodal probability distributions for deep ordinal classification. arXiv preprint arXiv:1705.05278 (2017)
22. Nishikawa, R.M., Comstock, C.E., Linver, M.N., Newstead, G.M., Sandhir, V., Schmidt, R.A.: Agreement between radiologists interpretations of screening mammograms. In: International Workshop on Digital Mammography, Springer (2016) 3–10
23. Salazar, A.J., Romero, J.A., Bernal, O.A., Moreno, A.P., Velasco, S.C.: Reliability of the bi-rads final assessment categories and management recommendations in a telemammography context. Journal of the American College of Radiology **14**(5) (2017) 686–692
24. Khan, M., Mohamed, S., Marlin, B., Murphy, K.: A stick-breaking likelihood for categorical data analysis with latent gaussian models. In: Artificial Intelligence and Statistics. (2012) 610–618
25. Sethuraman, J.: A constructive definition of dirichlet priors. Statistica sinica (1994) 639–650
26. Agresti, A.: An introduction to categorical data analysis. Volume 135. Wiley New York (1996)

27. Wan Kai, P.: Continuation-ratio model for categorical data: A gibbs sampling approach. In: Proceedings of the International MultiConference of Engineers and Computer Scientists. Volume 1. (2008)
28. Frigyik, B.A., Kapila, A., Gupta, M.R.: Introduction to the dirichlet distribution and related processes. Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006 (2010)
29. Gutiérrez, P.A., Tiňo, P., Hervás-Martínez, C.: Ordinal regression neural networks based on concentric hyperspheres. Neural Networks **59** (2014) 51–60
30. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin **70**(4) (1968) 213
31. 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. (2016) 770–778
32. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
33. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)