Leveraging Ordinal Information for Speech-based Depression Classification

Lishi Zuo, Man-Wai Mak

Dept. of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

lishi.zuo@connect.polyu.hk, man.wai.mak@polyu.edu.hk

Abstract

While depression is inherently ordinal, much of the previous work in depression detection oversimplifies the problem by treating it as a binary classification problem, ignoring the subtle variations and the order in depression severity. We propose creating a latent space that contains ordinal information via an ordinal loss to benefit the learning of depression classification. Specifically, we define K thresholds for the depression scores, thereby creating a series of binary classification tasks on different levels of depression (e.g., mild vs. non-mild). The ordinal loss allows the model to capture the relationships between these levels on top of the binary classification task. Our approach outperforms current state-of-the-art depression detection methods, highlighting the importance of considering the inherent ordinal nature of depression severity.

Index Terms: Speech-based depression detection, ordinal regression, ordinal classification

1. Introduction

Depression is a complex mental health condition that exists on a continuum, ranging from mild to severe. Prior work has shown that speech-based depression data is ordinal [1]. However, most detection models oversimplify this spectrum by reducing depression to binary classification—classifying individuals as either "depressed" or "non-depressed". These models binarize some standard depression scales, such as the eight-item patient health questionnaire depression scale (PHQ-8) [2], to determine whether a person suffers from depression or not. Specifically, individuals with PHQ-8 scores $s \geq T^{\text{gt}}$ are diagnosed as depressed, where T^{gt} is a predefined threshold in the depression scale. The binary simplification ignores the subtle variations in depression severity, as the model learns to distinguish between just two categories without considering intermediate levels.

Incorporating ordinal information for depression classification can be useful. To the authors' best knowledge, though previous research has applied ordinal regression for predicting depression severity [3, 4], ordinal information has not been utilized in depression classification. In this work, we propose an ordinal-dependent framework, which leverages ordinal data to equip depression classification models with the ability to *perceive* depression as a continuous spectrum. We refer to the resulting depression classification model as **Ordinal-Dep**. Rather than deciding if a patient is depressed or not by comparing a score with the predefined threshold T^{gt}, we introduce K additional thresholds $\{T^k\}_{k=1}^K$ and incorporate an auxiliary ordinal loss to complement the classification loss. This multi-threshold strategy enables the model to learn a latent representation $\boldsymbol{z}$ that captures the ordinal relationships between different severity levels (e.g., mild *vs.* non-mild). By learning representations that are effective across multiple thresholds instead of being confined to a single predefined threshold T^{gt}, the model can mitigate overfitting and better generalize to real-world variations in depression tasks.

In summary, our contributions are as follows.

1. To the authors' best knowledge, we are the first to incorporate ordinal information for speech-based depression detection.
2. We propose a framework that promotes the learning of latent representations with ordinal information using deep learning-based ordinal regression for speech-based depression classification. We empirically showed that incorporating ordinal information into a depression model can greatly enhance its performance, effectively mitigating the overfitting problem and improving generalization.

2. Related Work

2.1. Speech-based Depression Detection

Depression is a common and serious mental health disorder that leads to significant social, psychological, and economic consequences.[1] Speech is a valuable biomarker for detecting depression, as depressed individuals often exhibit slower speech, lower pitch, and reduced prosody variation, which manifest in acoustic features like speech rate, tone, pitch, and pauses [5, 6, 7]. Previous deep learning-based methods for speech-based depression detection primarily focus on model structure design [8, 9, 10], addressing bias and overfitting in low-resource settings [11, 12, 13], and protecting patient privacy [14]. However, these methods overlook the ordinal nature of depression. This work aims to leverage ordinal information for speech-based depression classification.

2.2. Ordinal Regression

Ordinal regression is useful for handling situations where data has a natural order, e.g., age prediction [15, 16], affective rating [17], and monocular depth estimation [18, 19]. One of the foundational approaches in ordinal regression is the proportional odds model, introduced by McCullagh [20], which relies on a strong statistical assumption that the odds of being in a higher category follow a proportional relationship across all thresholds. Alternatively, ordinal regression can be reformulated as a series of binary classification problems using multiple thresholds. This approach has been applied using perceptrons [21] and support vector machines [22]. For

This work was supported by RGC of Hong Kong SAR under Grant No. 15228223.

[1] https://www.who.int/news-room/fact-sheets/detail/depression

deep learning, a common technique is to transform ordinal regression into multiple binary classifications, where classifiers are learned to predict whether the input belongs to a particular category or falls within a given range [15, 16]. This transformation allows neural networks to leverage their capacity for high-dimensional feature learning while preserving the ordinal nature of the output. This work leverages deep-learning-based ordinal regression similar to [15, 16] for high-dimensional speech data for the depression classification task.

2.3. Ordinal Regression for Speech-based Depression

Speech-based depression data has been empirically ordinal [1]. Previous work shows that considering the ordinal relationships in the predicted scores while predicting the depression score can improve the performance of depression severity prediction [3, 4]. These approaches are based on traditional machine learning methods, such as the ranking support vector machine [3] and ordinal logistic regression [4], to predict depression scores. In this work, we brought the same intuition that accounts for the ordinal nature of depression data to deep-learning-based depression detection.

3. Method

Let $\mathcal{D} = \{(\boldsymbol{X}_i, y_i^{gt}, s_i)\}_{i=1}^N$ be the training set containing N segments/samples. Here, $\boldsymbol{X}_i$ comprises frame-based input feature vectors of the i-th segment, y_i^{gt} is the binary depression label (depressed/non-depressed), and s_i is the depression score indicating the severity of depression.

The overall framework is illustrated in Figure 1. The latent representation $\boldsymbol{z}$ is extracted by the encoder f, i.e., $\boldsymbol{z} = f(\boldsymbol{X})$. The depression classification head is denoted by e_1, while the depression ordinal head consists of $g(\cdot)$ and $e_2(\cdot)$. o^k represents the output from the k-th node of e_2. For implementation details, refer to Table 1.

3.1. Objectives

Depression detection is a binary classification problem, where the ground truth label y^{gt} is determined based on whether the ground truth depression score (e.g., PHQ-8) is higher than a predefined threshold T^{gt}. Specifically, the depression score s_i is compared to this threshold to assign the class label, i.e., $y_i^{gt} = \mathbb{1}(s_i \geq T^{gt})$. Given the ground truth label y^{gt}, the depression detection model can be trained:

$$\begin{aligned}\mathcal{L}_{cls} &= \mathbb{E}_{(\boldsymbol{X},y^{gt})}[w_1 y^{gt} \log p(s \geq T^{gt}|\boldsymbol{X}) \\ &\quad + w_0(1-y^{gt}) \log p(s < T^{gt}|\boldsymbol{X})] \\ &= \mathbb{E}_{(\boldsymbol{X},y^{gt})}[w_1 y^{gt} \log p(y=1|\boldsymbol{X}) \\ &\quad + w_0(1-y^{gt}) \log p(y=0|\boldsymbol{X})],\end{aligned} \tag{1}$$

where $p(s \geq T^{gt}|\boldsymbol{X})$ and $p(s < T^{gt}|\boldsymbol{X})$ are the classification model's outputs, and w_1 and w_0 denote the class weights for the depressed and healthy classes, respectively. The class weights are computed using the formula: $w_c = \frac{N}{N_c}$, where N_c is the number of segments in Class c and N is the number of training samples.

The depression classification loss $\mathcal{L}_{cls}$ ignores the ordinal nature of depression by treating depressed and non-depressed samples as distinct classes, potentially failing to capture the true underlying structure of the data. To better capture the ordinal relationships in the depression data, we propose incorporating additional thresholds and introducing an ordinal loss function, $\mathcal{L}_{\text{ord}}$, which encourages the model to consider the depression severity. Given $s \in [s_{\min}, s_{\max}]$, we manually define K' thresholds. These thresholds $T^{k'}$'s are uniformly spaced and are given by

$$T^{k'} = s_{\min} + k' \cdot \frac{s_{\max} - s_{\min}}{K'+1}, \quad k' = 1, \ldots, K'. \tag{2}$$

Given a set of thresholds $\{T^{k'}\}$, we define datasets $\mathcal{D}^{k'} = \{(\boldsymbol{X}_i, y_i^{k'}, s_i)\}_{i=1}^N$, where the label $y_i^{k'}$ for each threshold is determined by $y_i^{k'} = \mathbb{1}(s_i \geq T^{k'})$. In practice, depend on the data distribution in $\mathcal{D}$ and the defined $T^{k'}$, the dataset $\mathcal{D}^{k'}$ may be highly imbalanced, which can negatively impact model training. To enhance threshold effectiveness, we discard highly imbalanced datasets $\{\mathcal{D}^{k'}\}$ with $k' \in [1, K']$ for some k', and retain only those thresholds $T^{k'}$ for which the normalized positive class weight, $\eta^{k'} = \frac{w_1^{k'}}{w_0^{k'}}$, satisfies the condition $\frac{1}{l} < \eta^{k'} \leq l$, where $w_1^{k'}$ and $w_0^{k'}$ are the positive and negative class weights for the dataset $\mathcal{D}^{k'}$, respectively, and l is a hyperparameter that bounds the acceptable range of $\eta^{k'}$. In other words, we keep the datasets whose thresholds will not lead to a high imbalance between the positive and negative classes.

Among the filtered thresholds, K thresholds are randomly selected and kept, denoted as $\{T^k\}_{k=1}^K$. For notational simplicity, we denote the resulting datasets as $\mathcal{D}^k = \{(\boldsymbol{X}_i, y_i^k, s_i)\}_{i=1}^N$. The binary labels y_i^k are then used for binary classification tasks based on varying levels of depression severity, for example, predicting whether a sample corresponds to mild or non-mild depression. These multiple binary classification tasks are trained by minimizing $\mathcal{L}_{\text{ord}}$:

$$\mathcal{L}_{\text{ord}} = \frac{1}{K} \sum_k \mathbb{E}_{\mathcal{D}^k}\{(1 + (\eta^k - 1)y^k)(o^k - y^k)^2\}, \tag{3}$$

where $\eta^k = \frac{w_1^k}{w_0^k}$ is the normalized positive class weight of dataset $\mathcal{D}^k$ determined by threshold T^k. Here, $1 + (\eta^k - 1)y^k$ controls the class weight of the dataset $\mathcal{D}^k$. Specifically, for samples in $\mathcal{D}^k$ with $y^k = 1$, the corresponding class weights are η^k. For samples where $y^k = 0$, the class weight is set to 1. Utilizing η^k to address the class imbalance in $\mathcal{D}^k$ is crucial, particularly in imbalanced depression datasets $\mathcal{D}$, where $\mathcal{D}^k$ may exhibit even greater imbalance. Without η^k, the effectiveness of $\mathcal{L}_{\text{ord}}$ would be limited.

By incorporating additional thresholds, we transform the original binary classification task into multiple binary classification tasks, which encourages the model to learn a finer-grained latent representation, allowing it to not only predict the exact class but also capture ordinal information across different levels of depression. Consequently, the total loss is:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{cls}} + \beta \mathcal{L}_{\text{ord}}, \tag{4}$$

where α and β are the loss weights that balance the trade-off between $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{ord}}$.

3.2. Training

We dynamically adjusted the balance between ordinal loss $\mathcal{L}_{\text{ord}}$ and classification loss $\mathcal{L}_{\text{cls}}$, implementing a form of internal curriculum learning. Specifically, α was increased linearly from 0.5 to 1, while β was decreased from 2 to 1. Initially, the

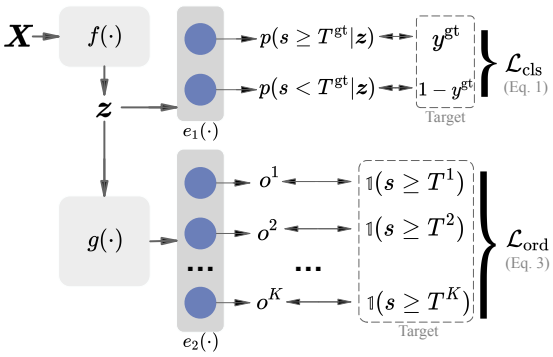

Figure 1: *The framework of Ordinal-Dep. Given the latent representation $\boldsymbol{z}$, $e_1(\cdot)$ outputs the posteriors of depression and non-depression. $e_2(\cdot)$ implements K binary classification tasks, with each output o^k compared with the ground truth label (0 or 1) of the corresponding threshold T^k.*

model focused on learning ordinal relationships, refining its understanding of severity progression before shifting to the depression classification task. This staged learning process ensures that the depression classification task benefits from a more structured latent space shaped by the ordinal loss.

4. Experimental Setup

4.1. Dataset

We evaluated Ordinal-Dep on the DAIC-WOZ corpus [23], a widely-used dataset for speech-based depression detection research. It comprises 189 individuals' interviews. The PHQ-8 scale was used to assess the participants, with scores $s \geq 10$ indicating depression ($T^{gt} = 10$). Figure 2 shows the score distribution across all speakers in the dataset. DAIC-WOZ contains training, development, and test subsets. We used the development set as the validation set and report the results on the test set. Long utterances were segmented into 3.84-second chunks, and frame-based Wav2Vec features [24] were extracted from these segments to serve as input for model training.

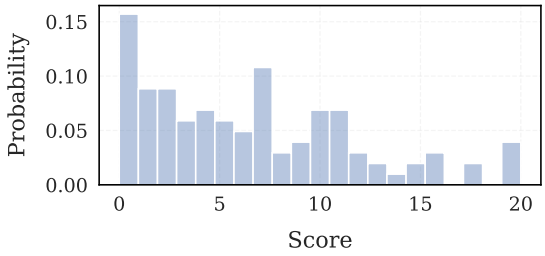

Figure 2: *The distribution of PHQ-8 scores across all speakers in DAIC-WOZ.*

4.2. Model Structure

The struture of the modules used in this study are shown in Table 1. All modules consist of fully connected layers. The hidden nodes of f and g employ the *tanh* activation function. The final layer of f is a statistics pooling layer that concatenates the mean, the standard deviation, and the mean of the first-order differences between successive feature frames across time.

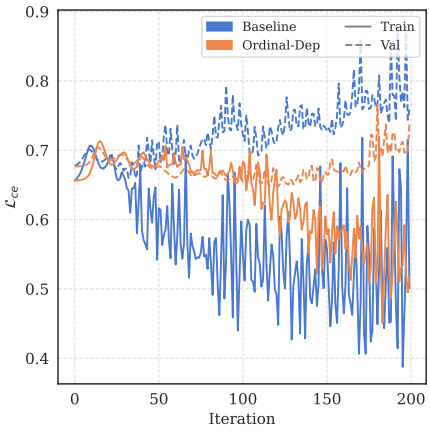

Figure 3: *Training and validation classification losses on the training set and the validation set.*

Table 1: *Implementation details of different modules in Ordinal-Dep. Here, dim_{in}, dim_{hidden}, and dim_{out} represent the input dimension, hidden dimension, and output dimension, respectively.*

Module	(dim_{in}, $\text{dim}_{\text{hidden}}$, dim_{out})
$f(\cdot)$	(512, 264, 116)
$g(\cdot)$	(48, -, 16)
$e_1(\cdot)$	(48, -, 2)
$e_2(\cdot)$	(16, -, 4)
$e_3(\cdot)$	(16, -, 1)

4.3. Training

In the experiments, we set $K' = 15$, $K = 5$, and $l = 2.5$ by default. The learning rate was initially warmed up from 0 to 3×10^{-3} over the first 10% of the total iterations, and then decayed following a cosine schedule for the remaining iterations. Each batch consists of 107 segments, with each segment randomly sampled from a distinct speaker within the training set. The model was trained for a total of 200 iterations, using the Adam optimizer.

4.4. Evaluation

We employed macro-averaged F1-score (MF1), F1-score for the positive class (F1-pos), and F1-score for the negative class (F1-neg) for evaluation. The classification decision for each subject (speaker) was based on the majority votes across the speech segments of their utterances. Each experiment was run five times with different initial states, and the average performance across these five runs was reported.

5. Results

5.1. Main Results

Table 2 (Rows 1-3 & 6) presents a comparison between the proposed Ordinal-Dep and state-of-the-art methods. The results show that Ordinal-Dep significantly outperforms other work that does not consider the ordinal nature of depression, highlighting the importance of introducing ordinal information to depression detection tasks.

In our ablation study (Rows 4-6 in Table 2), we also compare Ordinal-Dep with two frameworks: (1) Baseline:

Table 2: *Main results for the proposed Ordinal-Dep. Rows 1–3 present the results of state-of-the-art methods. Rows 4–5 present the results of the ablation study. "Regression-cls" represents a detector trained by minimizing a regression loss and a classification loss. "Ordinal-Dep" was trained by minimizing an ordinal regression loss and a classification loss.*

Row	Method	Dev			Test		
		F1-pos	F1-neg	MF1	F1-pos	F1-neg	MF1
1	**FVTC [25]**	0.440	0.810	0.625	0.410	0.780	0.595
2	**SIDD [12]**	0.745	0.866	0.805	0.481	0.721	0.601
3	**Spk-Emb [26]**	0.510	0.860	0.685	0.500	**0.820**	0.660
4	**Baseline (cls only)**	0.488 ± 0.033	0.833 ± 0.006	0.66 ± 0.020	0.300 ± 0.049	0.763 ± 0.017	0.532 ± 0.030
5	**Regression-cls**	0.644 ± 0.112	0.853 ± 0.014	0.749 ± 0.063	0.400 ± 0.112	0.754 ± 0.025	0.576 ± 0.052
6	**Ordinal-Dep**	0.690 ± 0.012	0.852 ± 0.002	0.771 ± 0.005	$\mathbf{0.580 \pm 0.025}$	0.794 ± 0.011	$\mathbf{0.690 \pm 0.011}$

model trained with depression classification loss only; and (2) Regression-cls: model trained with depression classification loss $\mathcal{L}_{\text{cls}}$ and regression loss $\mathcal{L}_{\text{reg}}$. Note $\mathcal{L}_{\text{reg}} = \mathbb{E}_{(X,s)}(s - e_3(g(\boldsymbol{z})))^2$, where $e_3(\cdot)$ is the output of the regression head.

The experimental results in Table 2 demonstrate that **incorporating ordinal information enhances the performance of depression classification**. In particular, Rows 4 & 6 show that Ordinal-Dep largely outperforms the baseline, highlighting the effectiveness of the proposed framework in leveraging ordinal information. While Regression-cls also utilizes score information and seeks to learn a finer-grained latent representation through mean squared errors, it does not explicitly account for ordinal relationships. As shown in Rows 5 & 6, Ordinal-Dep outperforms Regression-cls, demonstrating its superiority in utilizing score information by emphasizing depression as ordinal data. We conducted Mann-Whitney U tests [27] to compare the proposed Ordinal-Dep with both the Baseline and Regression-cls methods. Figure 4 shows that Ordinal-Dep demonstrates statistically significant superiority over the two ablated methods ($p < 0.05$).

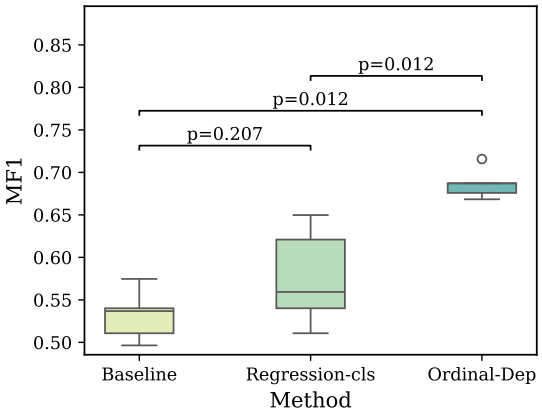

Figure 4: *Mann-Whitney U tests comparing the proposed Ordinal-Dep with Baseline and Regression-cls.*

We further found that the proposed **Ordinal-Dep significantly mitigates the overfitting problem** in the small DAIC-WOZ dataset. As illustrated in Figure 3, the training and validation classification losses for both Ordinal-Dep and its baseline are compared on the training and development sets of DAIC-WOZ. While the baseline model exhibits a large gap between training and validation loss, leading to rapid overfitting, Ordinal-Dep maintains a much smaller gap and ultimately achieves a lower validation loss. This improvement is likely attributed to the use of ordinal loss, which ensures that the model is trained not only to predict the exact class but also to respect the ordinal nature of depression, mitigating the risk of overfitting and potentially improving robustness on unseen data.

Additionally, **Ordinal-Dep stabilizes the training of $\mathcal{L}_{\text{cls}}$** (see Figure 3), likely due to the staged training process, which allows $\mathcal{L}_{\text{cls}}$ to begin learning from the structured latent representation initialized by $\mathcal{L}_{\text{ord}}$ (see Section 3.2).

5.2. Effect of K

Table 3 presents the effect of varying the number of thresholds, K, on the performance of the depression classification model. While increasing K is theoretically expected to improve performance by providing a finer-grained ordinal structure, our results indicate that $K = 5$ yields the best performance. Surprisingly, increasing K to 7 results in a decline in performance. We hypothesize that this may stem from variations in depression severity within different segments of speech from the same individual, suggesting that excessively granular classification levels may introduce bias. This, in turn, could negatively impact the model's accuracy. Moreover, managing a larger number of thresholds during training may induce instability, further complicating the learning process.

Table 3: *Effect of K on the performance of Ordinal-Dep on DAIC-WOZ.*

K	Dev			Test		
	F1-pos	F1-neg	MF1	F1-pos	F1-neg	MF1
2	0.557	0.839	0.698	0.450	0.784	0.617
3	0.689	0.846	0.768	0.563	0.774	0.668
5	0.690	0.852	0.771	**0.580**	**0.794**	**0.690**
7	0.636	0.779	0.707	0.514	0.746	0.630

6. Conclusions

In this paper, we highlight the importance of considering the ordinal nature of depression data in speech-based depression detection. We propose a framework that aims to create a latent space that contains ordinal information via an ordinal loss to benefit the learning of the depression detection task. By using the ordinal loss, the model is trained not only to predict the exact class but also to respect the ordinal nature of depression. The results show that the proposed method significantly outperforms state-of-the-art methods and effectively reduces the overfitting problem in the DAIC-WOZ dataset.

7. References

[1] S. Jayawardena, J. Epps, and Z. Huang, "How ordinal are your data?" in *Interspeech*, 2020, pp. 1853–1857.

[2] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[3] S. Jayawardena, J. Epps, and E. Ambikairajah, "Support vector ordinal regression for depression severity prediction," in *IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, 2018.

[4] S. Jayawardena, J. Epps, and E. Ambikairajah, "Ordinal logistic regression with partial proportional odds for depression prediction," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 563–577, 2020.

[5] J. K. Darby, N. Simmons, and P. A. Berger, "Speech and voice parameters of depression: A pilot study," *Journal of Communication Disorders*, vol. 17, no. 2, pp. 75–85, 1984.

[6] K. R. Scherer, "Vocal affect expression: A review and a model for future research." *Psychological Bulletin*, vol. 99, no. 2, p. 143, 1986.

[7] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *Journal of Psychiatric Research*, vol. 27, no. 3, pp. 309–319, 1993.

[8] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews." in *Proc. Interspeech*, 2018, pp. 1716–1720.

[9] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: An emotional audio-textual corpus and a GRU/BiLSTM-based model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6247–6251.

[10] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 35–42.

[11] A. Bailey and M. D. Plumbley, "Gender bias in depression detection using audio features," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 596–600.

[12] L. Zuo and M.-W. Mak, "Avoiding dominance of speaker features in speech-based depression detection," *Pattern Recognition Letters*, vol. 173, pp. 50–56, 2023.

[13] L. Zuo, M.-W. Mak, and Y. Tu, "Promoting independence of depression and speaker features for speaker disentanglement in speech-based depression detection," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 191–10 195.

[14] V. Ravi, J. Wang, J. Flint, and A. Alwan, "Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement," *Computer speech & language*, vol. 86, p. 101605, 2024.

[15] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognition Letters*, vol. 140, pp. 325–331, 2020.

[16] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4920–4928.

[17] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Transactions on Affective Computing*, p. 314–326, 2014.

[18] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.

[19] R. Diaz and A. Marathe, "Soft labels for ordinal regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4738–4747.

[20] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 42, no. 2, pp. 109–127, 1980.

[21] K. Crammer and Y. Singer, "Pranking with ranking," *Advances in Neural Information Processing Systems*, vol. 14, 2001.

[22] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *Proc. International Conference on Artificial Neural Networks*, vol. 1, 1999, pp. 97–102.

[23] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proc. the 9th International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 3123–3128.

[24] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469.

[25] Z. Huang, J. Epps, and D. Joachim, "Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6549–6553.

[26] S. H. Dumpala, S. Rodriguez, S. Rempel, M. Sajjadian, R. Uher, and S. Oore, "Detecting depression with a temporal context of speaker embeddings," *Proc. AAAI SAS*, 2022.

[27] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, pp. 50–60, 1947.