EVALUATION MEASURES FOR DEPRESSION PREDICTION AND AFFECTIVE COMPUTING

Sadari Jayawardena[1], Julien Epps[1,2], Eliathamby Ambikairajah[1,2]

[1]School of Electrical Engineering and Telecommunications, UNSW, Sydney
[2]Data61, CSIRO, Australia
s.jayawardena@unsw.edu.au, j.epps@unsw.edu.au

ABSTRACT

A variety of evaluation measures are being used to validate systems in depression prediction and affective computing. Among them, the most common measures focus on the error between the ground truth and predictions. However, when the ground truth is ordinal such as in psychiatric scores, ranking information is more important than the actual error. Therefore, this study systematically analyses the properties of classification, error-based and ranking measures particularly using classification accuracy, root mean square error (RMSE) and Spearman rank correlation coefficient, with the aim of identifying suitable measures for evaluating depression prediction and affective computing. For the purpose of analysis, we employed both synthetic data and real depression prediction systems evaluated with the AVEC2017 depression corpus. Outcomes of the experiments suggest that RMSE and classification accuracy, which are frequently used, are not sensitive to ordering and that rank correlation measures are more appropriate for depression prediction, which is an ordinal problem.

Index Terms— evaluation measures, ordinal regression, depression prediction, affective computing

1. INTRODUCTION

Depression is one of a number of mental disorders that collectively impose a high socio-economic burden on individuals. The cost of depression not only includes medical expenses but also the cost due to reduced working capacity and sometimes loss of life. Like many subjectively assessed quantities, depression severity is indicated using scales (such as PHQ-8 [1], BDI-II [2]), which are typically clinician-rated or self-rated. These scales are ordinal in nature, rather than numerical or categorical, because a depression score of 4 does not indicate double the severity of a score of 2 [3]. In this respect, representation of depression severity is similar to a great many other quantities in behavioural and affective computing, for example arousal, valence, dominance, cognitive load, level of interest, stress, anger etc.

Evaluation is an integral step in the development process for understanding the validity of classification and prediction systems. Numerous evaluation measures can be found in the literature for depression prediction. These measures can be categorized into three groups: classification, regression and ranking measures. By far the most commonly used measures in the field are from the former two groups, while less attention has been paid to ordinal performance evaluation. Each evaluation measure has its own strengths and weaknesses. Using inappropriate evaluation measures could result in poor system design choices, and therefore should be chosen with a full understanding of what information can be reflected about the underlying problem [4, 5].

With a view to guiding the choice of evaluation measures, this paper investigates the properties of evaluation measures for classification, regression and ordinal regression problems, with a particular focus on speech-based depression assessment because it has been evaluated in multiple different ways to date.

2. RELATION TO PRIOR WORK

Given the ordinal structure of the depression prediction problem, ordinal regression, also referred to as ordinal classification, is defined for dependent variables with ordered sequences. Define an input feature vector $x_i \in X$ and dependent variable $y_i \in Y$, with $Y = \{y_1 \prec y_2 \prec \cdots \prec y_K\}$, where K is the maximum value in the target scale. The goal in ordinal regression is to learn a mapping function $f: X \to Y$. Machine learning provides several models for the problem of ordinal regression including ordinal logistic regression [6, 7], support vector ordinal regression [8, 9], ordinal Gaussian processes [10], ordinal KDA [11] and preference learning based ranking models (e.g. RankSVM [12], deep neural network models [13, 14]). There is an emerging tendency towards ordinal regression in affective computing considering the ordinal nature of the data [15-17]. However ordinal regression modelling is out of the scope of this paper. An overview of evaluation measures found in depression prediction and affective computing is presented below.

The most common evaluation measure for classification is accuracy. In the literature, it is computed in one of three ways: (i) as the average of per class accuracy, also termed macro-average (ii) as the mean of accumulated true positives (micro-average) [4] (iii) *Average Weighted Accuracy (AWA)*: average accuracy weighted by class size. *Unweighted Average Recall (UAR)* is an extension of binary recall for multiclass classification and has been used in both depression prediction and affective computing [18, 19]:

$$UAR = \frac{\sum_{i=1}^{C} recall_i}{C}, \quad (1)$$

where C is the total number of classes. The kappa statistic [20] is a measure of inter-observer reliability and is given by:

$$k = \frac{P_o - P_e}{1 - P_e} \quad (2)$$

Kappa depends on two parameters: observed agreement P_o and expected agreement P_e. k = 0 is interpreted as chance agreement and positive and negative values indicate better and poorer than chance agreement respectively. In [21], kappa has been used to evaluate 3-class depression classification. The main limitation in using the above-mentioned classification measures for ordinal regression is their insensitivity to ordering. Classification measures

penalize all misclassifications equally, however measures for ordinal regression should be able to quantify the severity of the error. *Weighted kappa* [22] addresses this limitation by introducing weights proportional to the degree of disagreement. The weights influence the final measure, however deciding on the weights is not straightforward [23] since the metric distance is not defined in ordinal regression.

In regression measures, error is often calculated based on numerical distance. Two popular measures are *Mean Squared Error (MSE)* and *Mean Absolute Error (MAE)*:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (3)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (4)$$

where y_i and $\hat{y}_i$ are the ground truth and predicted values for the i^{th} input vector and N is the size of the test sample. To date, RMSE (Root MSE) and MAE are the basic evaluation measures in depression prediction [24] even though the metric distance between depression scores is meaningless. In [25], authors propose macro-averaged measures for ordinal regression. Macro-averaging helps to compensate for the bias due to data imbalance. Nevertheless, the above error-based measures are limited by their scale dependence: MAE and RMSE have the same units as the dependent variable, hence comparison across problems with different scales is not feasible. Furthermore, there is no absolute standard for a good measure value; comparison with a baseline is always necessary. Normalizing would resolve the above issues to some extent, e.g. *Normalized RMSE* [26], however NRMSE is not in common use in affective computing.

Correlation coefficients are used to assess the strength and direction of associations between pairs of variables and have also been used as measures to compare ground truth and predicted values. Unlike error-based measures, they are bounded within [-1, +1]. The most widely used correlation measure is *Pearson Correlation Coefficient* or *Pearson's r*:

$$r = \frac{\sum_{i=1}^{N}(y_i - \mu_Y)(\hat{y}_i - \mu_{\hat{Y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \mu_Y)^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \mu_{\hat{Y}})^2}} \quad (5)$$

where μ_Y and $\mu_{\hat{Y}}$ denote the means of ground truth and predictions. Pearson's r is defined for continuous, normally distributed variables and based on the assumption that the relationship among variables is linear. *Concordance Correlation Coefficient (CCC)* is normalized Pearson's r with the mean and variance of the two variables to penalize bias and scale variance between variables [27], and has often been employed for continuous emotion prediction:

$$CCC = \frac{2r\sigma_Y\sigma_{\hat{Y}}}{{\sigma_Y}^2 + {\sigma_{\hat{Y}}}^2 + (\mu_Y - \mu_{\hat{Y}})^2} \quad (6)$$

where μ_Y, σ_Y and $\mu_{\hat{Y}}, \sigma_{\hat{Y}}$ represent the means and standard deviations of ground truth (Y) and predictions ($\hat{Y}$). *Spearman's Rank Correlation Coefficient* (*Spearman's rho)* and *Kendall Rank Correlation Coefficient (Kendall's Tau)* [28] are rank correlation metrics, which are non-parametric and applicable to ordinal data. Furthermore, rank correlations are bias invariant, monotone invariant [29] and are robust to outliers [30]. Spearman's rho is the most widely used, and is based on ranking deviations (d):

$$\rho = 1 - \left(\frac{6\sum_{i=1}^{N} d^2}{N(N^2 - 1)}\right) \quad (7)$$

Spearman's rho frequently appears in the depression literature as a measurement of association between symptoms and depression scores [31, 32]. The expression for Kendall's tau is:

$$\tau = \frac{2\gamma}{N(N-1)} \quad (8)$$

where γ is the difference between the number of concordant and discordant pairs. A pair of instances is considered to be concordant only if predicted labels are in the same order with respect to ground truth labels, i.e.: $y_i > y_j \wedge \hat{y}_i > \hat{y}_j$ or $y_i < y_j \wedge \hat{y}_i < \hat{y}_j$. Since Kendall's tau is related to ranking order, it is more interpretable than Spearman's rho. Spearman's rho and Kendall's tau have been proposed to process ordinal labels in affective computing [33]. Apart from the above ranking measures, in [16] precision@k has been adapted for emotion recognition.

3. EVALUATION MEASURES IN SIMULATION

To test the appropriateness of the evaluation measures from Section 2 for ordinal problems, we conducted a series of experiments. To begin with we assessed chance-level system performance. This first experiment was motivational in nature: we took the AVEC2017 development dataset and computed (i) the chance-level RMSE for prediction (forcing the predictor output to the same PHQ-8 value for every test instance) and (ii) the chance-level accuracy for 5-class severity classification (forcing the classifier output to the same severity score value for every test instance). Depression categories were formed based on the definition of the PHQ-8 scale. Surprisingly, the lowest chance-level RMSE was 6.51, which is lower than the baseline result (6.74) for the database. Similarly, the highest classification accuracy (49%) was quite high for 5-class classification. Fig. 1 thus attests to the fact that rank information can be completely non-existent and yet RMSE or classification accuracy may appear competitive.

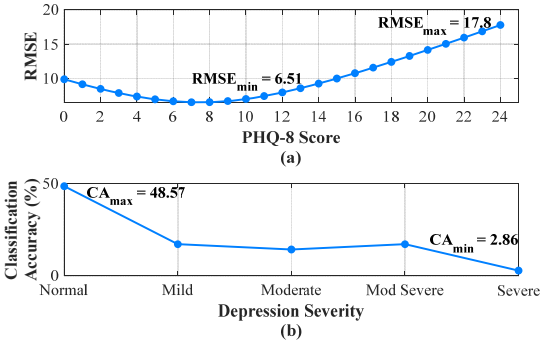

Fig. 1: (a) Chance-level *prediction RMSE* at each PHQ-8 score (b) chance-level *5-class classification accuracy (CA)* at each PHQ-8 severity level for AVEC2017 development dataset.

To probe this further, we observed the behaviour of rank correlations, this time under constant RMSE. We generated two random signals representing y_i and $\hat{y}_i$. Both signals were passed through a low pass filter with a cutoff frequency that was varied to modify the correlation between them. The RMSE between the signals was held constant by varying the filter gain while keeping the filter energy constant. Results are depicted in Fig. 2. The variation in Spearman's rho values demonstrates that RMSE is blind to correlation information, unlike ranking measures.

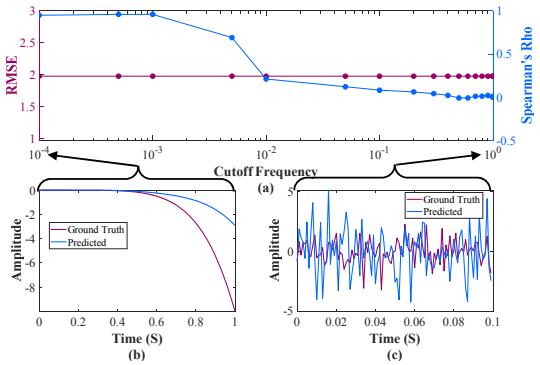

Fig 2: (a) Behaviour of Spearman's rho under constant RMSE. Cutoff frequency is in the range (10^{-4},1) where 1 corresponds to Nyquist frequency. (b) and (c) show the signals at cutoff frequencies of 10^{-4} (high correlation) and 1 (low correlation).

The next experiment was designed to analyse the performance of both RMSE and Spearman's rho with variation in correlation and bias. Without loss of generality, a baseline signal y_i (ground truth) was generated as an average of multiple random signals of the same length (1000 samples). A test signal $\hat{y}_i$ (predictions) was generated by low pass filtering the baseline signal and then adding an offset in the range [0, 5]. Unlike in previous experiment, decorrelation of the two signals was systematically increased by passing only the test signal through the low pass filter. The association between the two signals under different correlation effects (cutoff frequencies) and bias conditions (offsets) is presented in Fig. 3.

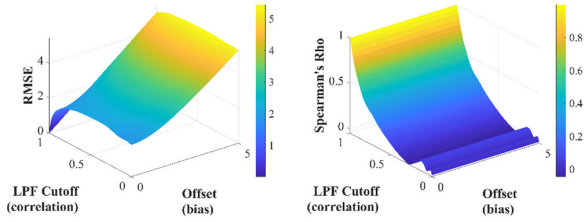

Fig. 3: Surface plot of (a) RMSE (b) Spearman's rho under varying cutoff-frequency and offset.

RMSE can be seen to monotonically increase with bias, whereas Spearman's rho is invariant to bias. Similar behaviour can be expected from both measures with monotone invariance, though it is not included in this study. Depression scores are indicative of relative order and their numerical values do not contain any other information. Therefore, bias invariance and monotone invariance are not critical as long as the correct ordering is preserved. On the other hand, since depression scores are upper and lower bounded, bias cannot exist with perfect positive correlation. RMSE appeared to have a non-injective relationship with correlation (controlled by LPF cutoff). Therefore, the same RMSE value can represent both high and low correlation and lower RMSE can be observed even when the signals have been designed with minimal correlation. These observations confirm the result from Fig. 2 that RMSE is insensitive to ordering information.

Finally, we compared classification accuracy against Spearman's rho to check whether classification accuracy is also inconsistent with ordering information. Herein, classification accuracy refers to micro-averaged accuracy. Ground truth labels were generated to simulate a uniformly distributed 5-class normal / mild / moderate / mod severe / severe classification problem under different severities of category swap errors. Classification accuracy and Spearman's rho were estimated under different error severity levels (Fig. 4). Classification accuracy clearly lacks ordering information, for example an accuracy of 36% can be observed at every error severity level. Classification accuracy is capable of identifying the existence of an error, but is less interpretable beyond that. In contrast, Spearman's rho shows monotonic negative correlation with error severity indicating its sensitivity to error magnitude.

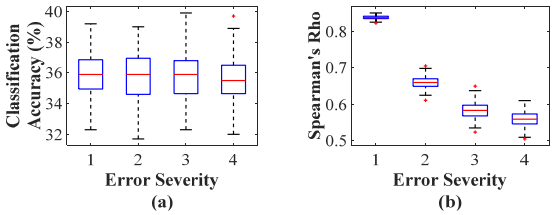

Fig. 4: Boxplot of (a) 5-class classification accuracy and (b) Spearman's rho as a function of error severity. Error severity represents the maximum allowed swap error (errors of 1 imply only adjacent category swaps).

4. EVALUATION MEASURES ON REAL DEPRESSION DATA

The preceding experiments were designed to provide insight into the properties of RMSE, classification accuracy and Spearman's rho for ordinal problems like depression prediction. However, it is most important to understand their behaviour on real depression data. For all experiments in this section, we used the AVEC2017 [24] dataset: models were trained using the training partition and tested on the development partition, unless otherwise stated.

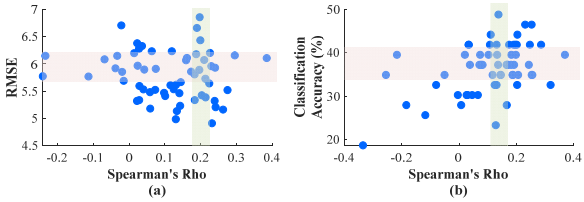

Fig. 5: Individual fold results for cross-validation on the AVEC2017 baseline system: (a) Random forest *regressor* with COVAREP features [34] (b) Random forest *classifier* with COVAREP features.

We performed cross-validation on the AVEC2017 dataset (training + development partitions), with a train-test split of 70:30. Different correlation values ranging from negative to postive can be observed for similar RMSE values (red colour band). Furthermore, different RMSE values can be observed with similar correlations (green colour band). Analogous observations can be made for the classification accuracy. These observations collectively agree with our results from synthetic data in Section 3: RMSE and classification accuracy both lack interpretability in terms of sensitivity to ranking information.

Fig. 6 presents the RMSE and Spearman's rho of multiple depression prediction systems. Note that the objective of this experiment was not to compare depression models but evaluation

measures, therefore default configurations were used when training the models. If RMSE were interchangeable with a ranking measure, then all points should occur on a monotonically decreasing line in Fig. 6. However, instead there are some major deviations, e.g. IS13+RFR and MFCC+GSR1 are not too different in RMSE but exhibit radically different correlation results.

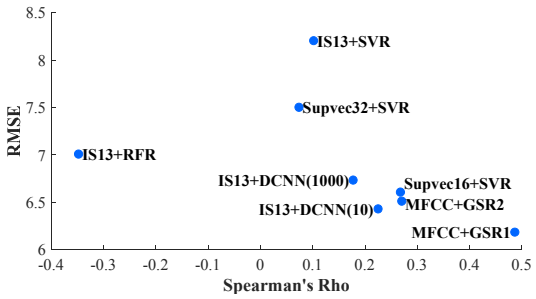

Fig. 6: Scatter plot of RMSE and Spearman's rho for multiple depression prediction systems. IS13: Interspeech 2013 feature set. SVR: Support Vector Regression. RFR: Random Forest Regressor. Supvec16 and Supvec32: GMM supervectors with 16 and 32 mixtures. DCNN(10) and DCNN(1000): Deep Convolutional Neural Network with 10 and 1000 epochs. GSR: Gaussian Staircase Regression (GSR1 and GSR2 have different partitioning in feature space).

In this study, classification accuracy, RMSE and Spearman's rho were used as representative measures of classification, regression and ranking measures respectively. Tables 1 and 2 present the correlations between all pairs of evaluation measures introduced in Section 2. The system used in these experiments was a DCNN with IS13 features. The architecture of the DCNN was the same as the system presented in [35]. To calculate multiple results from each evaluation measure, we considered all the combinations for 31-speaker subsets of the total 35 speakers in the development partition. Table 2 reports the results from 5-class classification.

Table 1: Pearson Correlation Coefficient between pairs of evaluation measures for *regression*. PCC = Pearson CC, SR = Spearman's rho, KT = Kendall's Tau.

	RMSE	NRMSE	MAE	PCC	SR	KT
RMSE	1	1	0.96	0.07	0	0
NRMSE	1	1	0.96	0.07	0	0
MAE	0.96	0.96	1	-0.01	-0.05	-0.04
PCC	0.07	0.07	-0.01	1	0.85	0.87
SR	0	0	-0.05	0.85	1	0.99
KT	0	0	-0.04	0.87	0.99	1

Neither RMSE, NRMSE nor MAE have an association with rank correlation measures. The insensitivity of these measures to ordering information may be the reason for this weak association (Fig 2). Pearson CC has a strong relationship with rank correlations, but it is not as strong as the association between Kendall's Tau and Spearman's Rho. Nevertheless, Pearson is a metric measure and not recommended for ordinal regression. Spearman's rho and Kendall's tau have similar underlying assumptions, therefore, very strong correlations can be observed.

Table 2: Pearson Correlation Coefficient between pairs of evaluation measures for *classification*. CA = Classification accuracy, WK = Weighted Kappa with linear weights

	CA	UAR	Kappa	WK	SR	KT
CA	1	0.7	0.68	0.69	0.53	0.52
UAR	0.7	1	0.74	0.57	0.31	0.3
Kappa	0.68	0.74	1	0.87	0.63	0.62
WK	0.69	0.57	0.87	1	0.92	0.91
SR	0.53	0.31	0.63	0.92	1	1
KT	0.52	0.3	0.62	0.91	1	1

Unlike classification accuracy and UAR, kappa and weighted kappa have a stronger association with rank correlations. Weighted kappa is equivalent to Pearson CC under certain conditions [22]. Therefore, a strong association can be expected.

5. CONCLUSION

This paper presents an overview of evaluation measures reported in depression prediction and affective computing literature to assess or select the best model for the problem. It is evident that each of these measures reflects an incomplete view of the actual error [5]. Therefore, selection of an evaluation measure must adhere to the expectations of the problem otherwise it is possible to end up with a non-optimal model.

Error-based measures (e.g. RMSE) are only lower bounded and hence without a baseline, their values have little meaning. Furthermore, error-based measures do not facilitate comparisons between depression corpora that have been labelled with different assessment scales. This is critical in depression and some other areas of affective computing, because databases have been labelled using different scales.

Based on experiments presented for synthetic data, the following two conclusions can be made: (i) both RMSE and classification accuracy are blind to ranking information. (ii) rank correlation measures are bias and monotone invariant. Bias and monotone invariance are not critical factors for psychiatric ratings. However, insensitivity to ranking information could lead to poor or incorrect model selection. In depression severity assessment, even when the error is low, if the correct ordering doesn't exist, it is hard to accept that model as a "good" model. In contrast, a model with high correlation but poor RMSE might still be acceptable, because of the ordinal nature of depression scores.

Using ranking measures for ordinal labels is not claimed as a novel [33]. Ordinal measures (including ranking measures) are the correct practise for ordinal data. However, until recently, ordering information has been quite widely ignored when evaluating models. Therefore, this work has highlighted the pitfalls of existing approaches and suggested that more elegant methods can be found in the ranking domain. These results provide encouragement to consider the careful choice of evaluation measures not only for automatic depression assessment systems, but for all types of affective computing problems for which the subjective ground truth is essentially ordinal in nature.

6. ACKNOWLEDGEMENTS

This work was partly supported by Australian Research Council Linkage Project LP160101360 and by Data61, CSIRO, Australia.

7. REFERENCES

[1] 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, pp. 163-173, 2009.

[2] A. T. Beck, R. A. Steer, and G. K. Brown, "Beck depression inventory-II," *San Antonio,* vol. 78, no. 2, pp. 490-498, 1996.

[3] N. Cummins, V. Sethu, J. Epps, J. R. Williamson, T. F. Quatieri, and J. Krajewski, "Generalized Two-Stage Rank Regression Framework for Depression Score Prediction from Speech," *IEEE Trans. on Affective Computing,* 2017.

[4] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management,* vol. 45, no. 4, pp. 427-437, 2009.

[5] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," *Geoscientific Model Development,* vol. 7, no. 3, pp. 1247-1250, 2014.

[6] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 109-142, 1980.

[7] A. Agresti, *Analysis of ordinal categorical data*. John Wiley & Sons, 2010.

[8] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Advances in Neural Information Processing Systems*, 2003, pp. 961-968.

[9] W. Chu and S. S. Keerthi, "Support vector ordinal regression," *Neural Computation,* vol. 19, no. 3, pp. 792-815, 2007.

[10] W. Chu and Z. Ghahramani, "Gaussian processes for ordinal regression," *Journal of Machine Learning Research,* vol. 6, no. Jul, pp. 1019-1041, 2005.

[11] B.-Y. Sun, J. Li, D. D. Wu, X.-M. Zhang, and W.-B. Li, "Kernel discriminant learning for ordinal regression," *IEEE Trans. on Knowledge and Data Engineering,* vol. 22, no. 6, pp. 906-910, 2010.

[12] R. Herbrich, T. Graepel, and K. Obermayer, *Regression models for ordinal data: A machine learning approach*. Citeseer, 1999.

[13] C. Burges *et al.*, "Learning to rank using gradient descent," in *Proc. of the Intl. Conf. on Machine Learning*, 2005, pp. 89-96: ACM.

[14] V. E. Farrugia, H. P. Martínez, and G. N. Yannakakis, "The preference learning toolbox," *arXiv preprint arXiv:1506.01709,* 2015.

[15] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!," *IEEE Trans. on Affective Computing,* vol. 5, no. 3, pp. 314-326, 2014.

[16] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5205-5209: IEEE.

[17] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *Intl. Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 248-255: IEEE.

[18] B. Schuller *et al.*, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load," in *Annual Conf. of the Intl. Speech Communication Association*, 2014.

[19] F. Ringeval *et al.*, "AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proc. of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 3-13: ACM.

[20] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement,* vol. 20, no. 1, pp. 37-46, 1960.

[21] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Trans. on Affective Computing,* vol. 4, no. 2, pp. 142-150, 2013.

[22] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bulletin,* vol. 70, no. 4, p. 213, 1968.

[23] M. Maclure and W. C. Willett, "Misinterpretation and misuse of the kappa statistic," *American Journal of Epidemiology,* vol. 126, no. 2, pp. 161-169, 1987.

[24] F. Ringeval *et al.*, "AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge," in *Proc. of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 3-9: ACM.

[25] S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation measures for ordinal regression," in *Intl. Conf. on Intelligent Systems Design and Applications*, 2009, pp. 283-287: IEEE.

[26] J. R. Cheema, "Some general guidelines for choosing missing data handling methods in educational research," *Journal of Modern Applied Statistical Methods,* vol. 13, no. 2, p. 3, 2014.

[27] F. Weninger, F. Ringeval, E. Marchi, and B. W. Schuller, "Discriminatively Trained Recurrent Neural Networks for Continuous Dimensional Emotion Recognition from Audio," in *IJCAI*, 2016, pp. 2196-2202.

[28] M. G. Kendall, "A new measure of rank correlation," *Biometrika,* vol. 30, no. 1/2, pp. 81-93, 1938.

[29] W. Xu, Y. Hou, Y. Hung, and Y. Zou, "A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models," *Signal Processing,* vol. 93, no. 1, pp. 261-276, 2013.

[30] S. Rosset, C. Perlich, and B. Zadrozny, "Ranking-based evaluation of regression models," in *Intl. Conf. on Data Mining*, 2005, p. 8 pp.: IEEE.

[31] A. Arseniev-Koehler, S. Mozgai, and S. Scherer, "What type of happiness are you looking for?-A closer look at detecting mental health from language," in *Proc. of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018, pp. 1-12.

[32] B. L. Brody *et al.*, "Depression, visual acuity, comorbidity, and disability associated with age-related macular degeneration," *Ophthalmology,* vol. 108, no. 10, pp. 1893-1900, 2001.

[33] G. N. Yannakakis, R. Cowie, and C. Busso, "The Ordinal Nature of Emotions: An Emerging Approach," *IEEE Trans. on Affective Computing,* 2018.

[34] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 960-964: IEEE.

[35] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal Measurement of Depression Using Deep Learning Models," in *Proc. of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 53-59: ACM.