Ordinal Logistic Regression with Partial Proportional Odds for Depression Prediction

Sadari Jayawardena, Julien Epps, Eliathamby Ambikairajah

Abstract— Like many psychological scales, depression scales are ordinal in nature. Depression prediction from behavioural signals has so far been posed either as classification or regression problems. However, these naive approaches have fundamental issues because they are not focused on ranking, unlike ordinal regression, which is the most appropriate approach. Ordinal regression to date has comparatively few methods when compared with other branches in machine learning, and its usage is limited to specific research domains. Ordinal logistic regression (OLR) is one such method, which is an extension for ordinal data of the well-known logistic regression, but is not familiar in speech processing, affective computing or depression prediction. The primary aim of this study is to investigate proportionality structures and model selection for the design of ordinal regression systems within the logistic regression framework. A new greedy based algorithm for partial proportional odds model selection (GREP) is proposed that allows the parsimonious design of effective ordinal logistic regression models, which avoids an exhaustive search and outperforms model selection using the Brant test. Evaluations on the DAIC-WOZ and AViD depression corpora show that OLR models exploiting GREP can outperform two competitive baseline systems (GSR and CNN), in terms of both RMSE and Spearman correlation.

Index Terms— Depression Prediction, Logistic Regression, Partial Proportional Odds, Model Selection

1 INTRODUCTION

MAJOR Depressive Disorder (MDD), referred to simply as depression, is highly prevalent in society. More than 300 million people, representing 4.4% of the world's population, were diagnosed with depression in 2015 [1], and these numbers are growing. If left untreated, depression can lead to serious consequences including suicide, self-harm, unhealthy lifestyles and disruptions to the activities of normal daily life. To prevent such consequences, early diagnosis and treatment is critical.

Similarly to other psychiatric disorders, depression is diagnosed through clinical interviews and mental state examinations [2]. To standardise the diagnostic process, various assessment tools have been developed and these tools consist of questionnaires covering a range of depression symptoms and could be either clinician-administered or self-evaluated. In the former approach, the questionnaire is completed by the clinician, whereas in the latter it is done by the patient themselves. The accuracy of the assessment of depression depends on two criteria: *reliability* and *validity* [3]. Reliability could be either interpersonal reliability, i.e. the level of agreement among assessors or intrapersonal reliability, i.e. the consistency over time. Validity is the *'correctness'*. The reliability and validity of the assessment greatly depends on clinician's knowledge and experience as well as on the patient's awareness and honesty. Considering the lack of reliability and validity in current diagnostic methods, machine learning based depression assessment that is objective and convenient has become an attractive prospect.

Speech has been identified as an objective measure that is sensitive to mental state changes. When a person is depressed, their speech typically becomes dull and monotonous [2]. Human speech production is a complex task involving cognitive effort and motor coordination of the articulators [4]. Any functional impairment in these systems could result in measurable effects on the speech signal [2], which can be quantified using methods from digital signal processing. Additionally, speech-based depression prediction is an inexpensive, non-intrusive and non-invasive approach.

To date, the problem of automatically assessing depression from speech has been addressed using classification or regression [5-8]. However, there is a fundamental issue with posing the problem either as classification or regression. Classification is for nominal response variables, where there is no relationship among classes: misclassification of an instance belonging to class *severe depression* to *normal* would be penalized equally to misclassifying the instance as *mild depression*, even when the former outcome is clearly more incorrect than the latter. Regression is applied to numerical response variables, typically in this setting to predict the depression severity as a numerical score (e.g. Beck Depression Inventory [9], Patient Health Questionnaire [10]). As mentioned above, the depression score is generated from a questionnaire evaluated on a Likert scale, hence the responses (sub-scores) are ordinal. The final depression score, which is the summation of numerically interpreted sub-scores, is neither strictly ordinal nor strictly numerical. For example, sub-scores in PHQ-8 are labels for frequency of occurrence (section 2.1) and hence aggregation of sub-scores would not produce a numerical variable (i.e. an aggregated depression score of 4 does not imply twice the severity of a depression score of 2 [11]). Yet they have a partial ordering among them. Thus, posing

- *Sadari Jayawardena is with the University of New South Wales, Sydney, Australia. E-mail: s.jayawardena@unsw.edu.au.*
- *Julien Epps is with the University of New South Wales, Sydney, Australia. E-mail: j.epps@unsw.edu.au.*

depression assessment as a regression problem is not sound mathematically, and more focus should be placed on ranking information. Hence, herein we focus on depression prediction using an ordinal regression approach. Ordinal regression is a branch of supervised learning where the dependent variable is ordinal categorical. It has resemblance to both classification and regression. However, unlike regression, the number of categories in the ordinal scale is finite and the ordinal relationships among the categories make ordinal regression distinct even from classification.

Generalised Linear Models (GLMs) are a family of models for different types of categorical dependent variables: dichotomous (e.g. logistic regression), polytomous and unordered (e.g. multinomial logistic regression), polytomous and ordered (e.g. ordinal logistic regression). The most frequently used ordinal logistic regression model is the Cumulative-Odds Model (COM) [12]. The COM has some conceptual similarities with the Gaussian Staircase Regression (GSR) model [7]. Instead of logits, GSR compares likelihoods of low and high partitions in the feature space. Based on its outstanding performance in both of the AVEC 2013 [7] and 2014 [13] challenges, GSR is a reference model for depression prediction, and one whose structure of pairwise likelihood ratios acknowledges to some extent the ordinality of the problem. The impressive modelling capacity of GSR motivated us to study COM, as well as its alternative model choices offered by the ordinal logistic regression framework.

Unlike *logistic regression*, the application of *ordinal logistic regression* is minimal in the machine learning literature, despite being one of the standard ordinal models. Therefore, this study aims to develop guidance for ordinal logistic regression system design for non-trivial amounts of data, while exploring model structure choices (Section 3.1) and the properties of proportionality structures (Section 3.2-3.4). Furthermore, we propose a greedy algorithm to automatically determine the proportionality structure for ordinal logistic regression framework, to avoid the computationally intensive task of parsimonious and effective design of the model structure.

2 RELATED WORK

2.1 Automatic Depression Prediction

In the literature, diverse modelling methodologies have been examined with respect to depression prediction. Depression prediction has three forms: 1) detecting the presence of depression: depressed or non-depressed 2) depression severity prediction: e.g. *Normal, Mild, Moderate, Severe* 3) depression score prediction, i.e. predicting the correct numerical score on a depression scale [2]. So far, these three forms have been solved as binary-classification, multi-class classification and regression respectively. The classification algorithms that appear frequently in the literature are Support Vector Machines (SVM) [6, 14, 15], Gaussian Mixture Models (GMM) [15, 16], decision trees [17, 18] and Deep Neural Networks (DNN) [19]. Support Vector Regression (SVR) [20], decision trees [21], Random Forest Regression [22], Relevance Vector Machines (RVM) [23], Gaussian processes [24], Extreme Learning Machine (ELM) [25] and DNN [26, 27] are few regression models in depression score prediction. It is of note that uncertainty inherent in the depression prediction models is usually ignored. Quantifying and incorporating uncertainty in model prediction to improve the system performance remains an open problem.

In the context of depression prediction, formants, pitch (F0) and spectral based measures are among many previously investigated acoustic features [2]. Spectral features are easy to compute and yield good performance [28]. MFCC is the most commonly-used spectral feature and its suitability for depression prediction has been assessed in previous studies [2]. Detailed spectral features including MFCC and Spectral Centroid Frequency (SCF) are susceptible to speaker and phonetic variability but still tend to perform better than broad spectral features such as energy slope and zero-crossing rate [29]. Based on the study in [30], it is reasonable to assume that detailed spectral features are well suited for the GSR model.

Brute-forced representations are produced by applying number of statistical functionals over different frame-level features either from single or multiple modalities. Depression affects people differently (symptoms may be heterogenous), i.e. patients do not share the same set of symptoms at the same severity level [31], which implies that depression could affect speech in many different ways. Therefore, in principle brute-forcing helps capture a diversity of speech effects induced by depression, and has been shown to be effective for prediction [11]. eGeMAPS [32] and COVAREP [33] are two such well-known brute-forced feature sets. Another advantage of brute-forced features is standardisation of feature extraction. Hence these features are common baseline feature sets in speech related applications. Brute-forced features typically provide a single vector for each utterance. When frame-level features are preferred LLDs without functionals can also be used instead [34]. Phone Log Likelihood Ratio (PLLR), a representation of frame-level phonetic probability, is a feature set that has been recently introduced to emotion and mental state prediction [34], while its application to language recognition [35] and speaker recognition [36] is already well established. Considering the promising performance improvements in other domains, it is of interest to also investigate PLLR for depression prediction.

BDI-II and PHQ-8 are two self-evaluated depression assessment tools. BDI-II [37] is the latest version of the original Beck Depression Index (BDI), which was developed to measure the intensity of depression in adolescents and adults. It evaluates 21 multiple-choice questions covering number of symptoms of depressive disorder. Intensity of each sub-symptom is rated on a 0-3 scale and hence the BDI scale range is 0 to 63. The eight-item Patient Health Questionnaire (PHQ-8) [38] comprises 8 questions to measure the frequency of occurrence of 8 depression symptoms. The rating for each question is from 0 to 3, where 0 and 3 represent *not at all* and *nearly every day* respectively. The total score range of PHQ-8 is 0 to 24.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3031300, IEEE Transactions on Affective Computing

2.2 Ordinal Regression

The problem of ordinal regression can be formulated as follows: for an input feature vector $x_i \in X$ and dependent variable $y_i \in Y$ s.t. $Y = \{y_1 \prec y_2 \prec \cdots \prec y_K\}$, the goal in ordinal regression is to learn the mapping function $f: X \rightarrow Y$. K is the number of ordinal categories. In terms of depression prediction, little work has been done on ordinal regression relative to classification and regression. Gaussian Staircase Regression or GSR [7] is a technique that takes some account of the ordinal nature of depression scores by calculating likelihood ratios from pairs of Gaussians – low and high – partitioned along the depression scale in a stepwise manner (hence the name). The LMLR (Log Mean Likelihood Ratio), logarithm of linear summation of likelihood ratios, each of which quantifies one of the pairwise low-high comparisons, are then collectively applied to predict depression scores, using linear regression. Motivated by the concept of pairwise comparison in the GSR model, a generalized framework for GSR was introduced in [11]. Both Weighted GSR and RVM-SR (Relevance Vector Machine Staircase Regression) make use of the concept of staircase regression to account for the ordinal nature of depression scores. Even though the essence of GSR is pairwise comparisons in feature space partitions, its output is a continuous value. Thus, GSR is not strictly an ordinal regression model.

Most ordinal regression models found in machine learning are either ranking models or redesigned models from multiclass classification. Ranking is a sub-branch of ordinal regression where the output is an ordered sequence of instances. In ranking, preference learning (PL) [39] is quite popular, which reduces the ranking problem to pairwise comparisons, i.e. for any given two instances A and B, learning whether A is preferred over B or not. The pairwise comparisons are then converted into ranks. A review of ordinal regression models is presented below.

PRank (Perceptron Ranking) [40] is a ranking algorithm in which a set of perceptron weights and bias are updated iteratively until ranking loss becomes zero or minimized, somewhat similar to *backpropagation*. RankSVM [41] is an extension of support vector framework for ranking which uses a loss function based on preference learning. Other preference based ranking models include Gaussian processes [42], neural networks (e.g. RankNet [43], PrefNet [44]), boosting algorithms (e.g. RankBoost [45]) and constraint classification [46]. A comparison of Naive Bayes ranking models is presented in [47]. The output of ranking algorithms is a ranking score, which is continuous. Hence, when using ranking models to solve ordinal problems, a mapping function is required to convert continuous ranking scores into ordinal categories. This mapping function is not explicit and has to be learnt during model fitting. Additionally, preference learning is computationally expensive due to pairwise comparisons and thus does not scale well for large datasets.

Multiclass classification models redesigned for ordinal regression include Gaussian processes [48], kernel discriminant learning [49], ELM [50, 51] and SVM [52], while [53] deployed nested binary classifiers. Redesigned classifiers for ordinal regression inherit the pros and cons from the original model. For example, ordinal GP suffers from poor scalability to large datasets, and ordinal SVMs and ANNs come with the burden of hyper-parameter selection. To the best of the authors' knowledge, except RankSVM [54], other ordinal regression models have not been investigated for depression prediction or related applications.

The family of GLMs provides a wide range of models for both classification and ordinal regression. Unlike the ordinal models presented above, GLM provides models that were originally proposed for ordinal regression. These models follow two assumptions: 1) the dependent variable Y follows a probability distribution such as Logistic, Normal, Poisson, and Cauchy; 2) the dependent variable Y is a discretised form of Y' which is a latent and continuous variable. The latent variable Y' is comparable to the ranking scores of a ranking model. The relationship between predictor variables $X \in \mathbb{R}^d$ and latent variable Y' is characterised as a linear regression model. The dependent variable Y and the latent variable Y' are related by a link function $G(\pi)$. The general equation for GLMs is given in (1). The link function is a monotonic mapping from [0, 1] to [-∞, +∞] which eliminates boundedness and makes the rest of the GLMs differ from linear regression. A widely used link function is the logit function $G(\pi) = log(\pi/(1-\pi))$ which is the inverse of logistic distribution. π is the probability of the success of a given event. Models presented in Section 3 use the logit link function.

$$G(\pi) = Y' = \alpha + \sum_{i=1}^{D} \beta_i X_i \tag{1}$$

GLMs have an equivalence to neural networks. The basic unit in a neural network, the neuron (or node), calculates the linear combination of the input feature vector using a weight vector and passes the output through a nonlinear activation function which is similar to the behaviour of GLMs. Neural networks consist of more than one such neuron spread across multiple layers, allowing repeated application of the neuron function, which makes the neural networks capable of forming non-linear complex relationships. However, such complex models are not necessary for every problem, especially when the dataset is not too large. In this regard GLMs, which are a simpler form of neural networks, can be more elegant.

Binary logistic regression or simply logistic regression is one of the popular models among GLMs. The application of logistic regression to speech processing is relatively less common compared with other binary classifiers. In early studies, logistic regression was rarely used as a prediction model, instead it has been used for fusion and Voice Activity Detection (VAD) [21, 55, 56]. Coefficients of logistic regression (linear component) represent a log-odds ratio for the corresponding predictor variable and hence act as an implicit indicator of importance of each subsystem. Logistic regression-based fusion has been applied in depression prediction as well [29, 57]. In [21], logistic regression was used as a VAD to identify voiced frames with a voicing threshold of 0.5. An application of logistic regression for depression prediction can be found in [58], where logistic regression was used to predict positive responders to depression treatment using two vocal prosody features: F0

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3031300, IEEE Transactions on Affective Computing

and speaker switch duration. In [59], ensemble logistic regression was used to estimate the presence of depression.

Logistic regression, rather than being a single method, comprises a hierarchy of models for ordered prediction problems, herein collectively referred to as Ordinal Logistic Regression (OLR). OLR is well-studied in the biomedical and epidemiology research communities because their studies are often related with ordinal outcomes such as severity of diseases (e.g. coronary heart disease, diabetes) and response levels for treatments. OLR is not yet studied thoroughly in speech-based mental state recognition, which is often challenging in terms of large dataset sizes, high feature dimensions and large amounts of unwanted variability. Even still, it is one of the few standard ordinal models which can directly predict ordinal categories through probabilities. This fact motivated us to perform a systematic analysis of ordinal logistic regression models.

2.3 Model Selection

The proportionality structure of ordinal logistic regression models (Section 3.2-3.4) offers multiple model configurations from which a design must be selected. The following is a review of model selection found in machine learning. Model selection can be characterised as finding the best model from a set of candidate models for a given data set. In order to identify the '*best*' model, an evaluation criterion or a benchmark is required. Evaluation criteria could be a goodness-of-fit criteria such as maximum-likelihood, an error-based measure such as Mean Absolute Error (MAE) or a correlation-based measure such as Spearman correlation. Some of these criteria tend to choose the most complex model and hence could result in overfitting [60]. In contrast, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) assess both likelihood and model complexity and hence discourage overfitting [60-62]. Later, Minimum Description Length (MDL) [63] was introduced, which states that the model with shortest description of data should be chosen [64].

In model selection, especially when there is a very large space of possible model configurations, the search strategy is critically important. A naive strategy is the exhaustive selection strategy, in which all the candidate models are evaluated to select the '*best*' model. However, this approach is not practical when the size of the candidate model pool is large. Forward, backward and stepwise (forward + backward) selection are three greedy algorithms found in subset selection. Forward selection starts with the '*null*' model and looks for the '*best*' model by adding one variable (which is the '*best*') at a time. In contrast, backward selection starts with the '*full*' model and removes one variable (which is the '*worst*') at a time. Stepwise selection is a combination of both forward and backward selection. It starts with the '*null*' model and takes a *forward* step. At each step, if there are any insignificant choices, stepwise selection would make a *backward* step. Greedy selection algorithms always use a greedy heuristic to make the locally best choice at each turn. Therefore, it is possible to converge into a locally optimal solution, but still greedy solutions are generally faster [65]. Hence, greedy algorithms are quite common in model selection literature. Greedy forward selection has been applied to find the optimum mixture number in GMM modelling in [66]. Feature selection is another domain where greedy selections are frequently utilised [67].

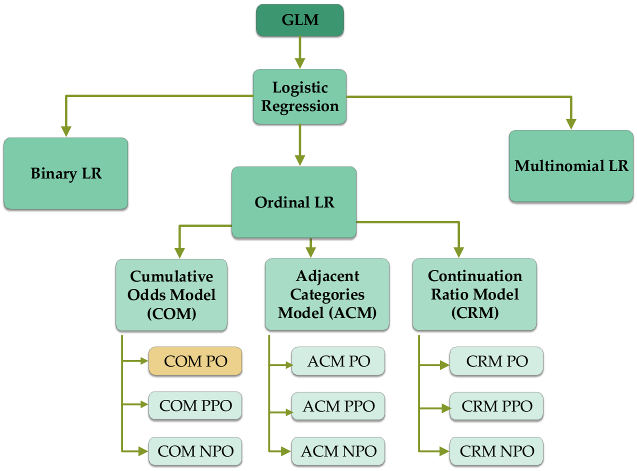

Fig. 1. Family Tree of Ordinal Logistic Regression. PO - Proportional Odds, PPO – Partial Proportional Odds, NPO – Non-Proportional Odds. COM PO, coloured in yellow, is the most common ordinal logistic regression model.

3 Ordinal Logistic Regression for Depression Score Prediction

Starting from a brief overview of logistic regression, this section describes the novel application of ordinal logistic regression to depression score prediction. Figure 1 summarises the hierarchy of the models discussed later in this paper.

3.1 Ordinal Logistic Regression Overview

Logistic regression uses the logit function, or sigmoid function, for the purpose of binary classification, which is defined as:

$$log\left(\frac{\pi}{1-\pi}\right) = \alpha + \sum_{i=1}^{D} \beta_i X_i \qquad (2)$$

where $\pi = P(Y = 1|X)$ represents the probability of being depressed given the acoustic feature vector X. Y specifies the two outcomes: depressed ($Y = 1$) or not ($Y = 0$). D is the number of predictors. Coefficients β_i are the log odds ratios for the corresponding predictor variable and α is known as the *cut-point* on the latent variable scale Y'. A maximum likelihood (ML) based method is currently the most generally employed approach for coefficient estimation [68].

3.1.1 Cumulative-Odds Model (COM)

When the dependent variable is polytomous, such as depression scores on a scale like PHQ-8, the probability of success can be conceived in multiple ways. The most common approach is by accumulating probabilities, hence the name Cumulative-Odds model:

$$log\left(\frac{\pi_k}{1-\pi_k}\right) = \alpha_k + \sum_{i=1}^{D} \beta_i X_i \qquad (3)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3031300, IEEE Transactions on Affective Computing

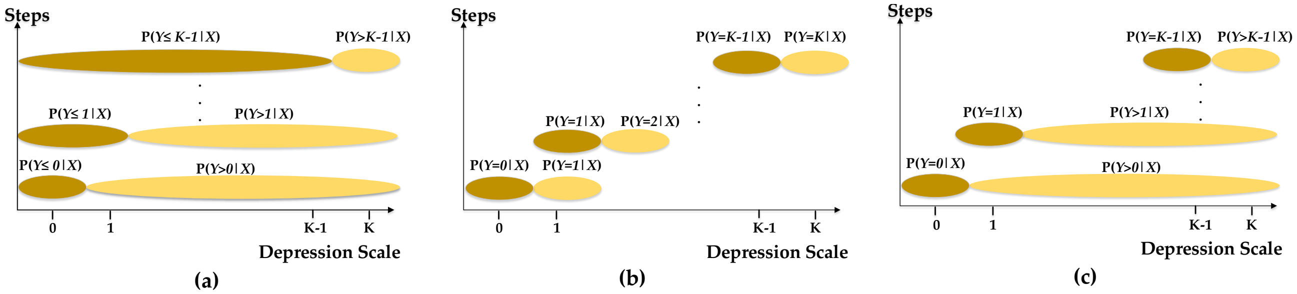

Fig. 2. Logit formulation in (a) Cumulative-Odds (COM) (b) Adjacent-Category (ACM) and (c) Continuation-Ratio (CRM) models. The three models accommodate three different pairwise comparisons along the depression scale which is ordinal in nature.

where $\pi_k = P(Y \leq k|X)$ represents the probability of a depression score being less than or equal to the score value k, where $k \in \{1, 2, \cdots, K-1\}$ and K is the maximum value in the target scale range. Intercept α_k satisfies the condition $\alpha_1 < \alpha_2 < \cdots < \alpha_{K-1}$. In general, $\alpha_0 = -\infty$ and $\alpha_K = +\infty$. It is assumed that $y = k$ only if $\alpha_{k-1} < y' \leq \alpha_k$, where Y' is the latent, continuous variable. Equation (3) can be conceptualised as a stepwise partitioning of the depression scale into dichotomous categories (Fig. 2(a)), analogous to the 'staircase' of Gaussians employed in GSR [7, 11, 30].

3.1.2 Adjacent-Category Model (ACM)

In the cumulative-odds model, the behaviour of the dependent variable is analysed by comparing the cumulative probabilities of dichotomous partitions. The Adjacent-Category model offers a different approach to analyse depression by considering only pairs of adjacent depression levels (Fig. 2(b)).

$$log\left(\frac{\pi_k}{\pi_{k+1}}\right) = \alpha_k + \sum_{i=1}^{D} \beta_i X_i \tag{4}$$

In (4), $\pi_k = P(Y = k|X)$ is the probability of depression score being k. Hence, this model does not use all data at every step. ACM is more appropriate when the categories of the response variable are substantive [69], such as when comparing the depression severity of multiple patients.

3.1.3 Continuation-Ratio Model (CRM)

The Continuation-Ratio model is used when the underlying process is sequential [70]. Since depression is progressive, i.e. a patient who is suffering from a certain depression level has passed through all the lower depression levels, depression prediction also can be posed using CRM. Similarly to ACM, CRM does not use the full dataset at each step: it ignores the data that are in lower categories than severity level k (Fig. 2(c)).

$$log\left(\frac{\pi_k}{\sum_{l=k+1}^{K} \pi_l}\right) = \alpha_k + \sum_{i=1}^{D} \beta_i X_i \tag{5}$$

In (5), $\pi_k = P(Y = k|X)$ is the probability of depression score being k, which is compared against the probability of depression score being in a higher level than k.

3.2 Proportional Odds (PO) Assumption

In (3), (4) and (5), only the intercepts (α_k) depend on the depression severity level k, but not the coefficients (β). This means that a common effect is assumed for predictor variables across all severity levels and this assumption is referred to as the *proportional odds assumption* or *assumption of parallelism*. The proportional-odds assumption offers the simplest model structure, and is referred to as a Proportional Odds (PO) model (Fig. 3(a)) with $D + K - 1$ model parameters.

3.3 Non-Proportional Odds (NPO)

Even though the proportional odds assumption allows easily interpretable models, it is not valid for every dataset, i.e. it may not be realistic to assume that depression symptoms will have a constant variation across all severity levels. For an example, the study in [5] illustrates the negative, non-monotonic relationship between Average Weighted Variance (AWV) and severity of depression. When the proportional odds assumption is not satisfied, a generalised model in which the relationship between predictor variables and dependent variable differ with the depression level k (Fig. 3(c)) can be fitted to the data. The non-proportional odds (NPO) structure for COM, ACM and CRM is given by (6), (7) and (8) respectively.

$$log\left(\frac{\pi_k}{1-\pi_k}\right) = \alpha_k + \sum_{i=1}^{D} \beta_{ik} X_i \tag{6}$$

$$log\left(\frac{\pi_k}{\pi_{k+1}}\right) = \alpha_k + \sum_{i=1}^{D} \beta_{ik} X_i \tag{7}$$

$$log\left(\frac{\pi_k}{\sum_{l=k+1}^{K} \pi_l}\right) = \alpha_k + \sum_{i=1}^{D} \beta_{ik} X_i \tag{8}$$

In the above model structures, the coefficients (β_{ik}) depend on each depression level k. Relaxation of the assumption of parallelism makes the model more complex by introducing significantly more parameters: NPO has $(D + 1) \times (K - 1)$ model parameters in contrast with $D + K - 1$ for PO. A drawback associated with NPO is that (as with all other machine learning methods) when the model is less parsimonious, a large dataset is required to estimate model coefficients, otherwise it could lead to overfitting.

3.4 Partial-Proportional Odds (PPO)

Partial Proportional Odds [71] (Fig. 3(b)) is proposed herein for depression score prediction, as a compromise between PO and NPO. PPO provides more modelling flexibility to trade-off between model parsimony and

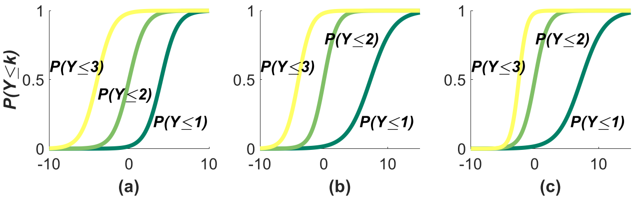

Fig. 3: Depiction of (a) Proportional Odds ($\beta_1 = \beta_2 = \beta_3$); (b) Partial Proportional Odds ($\beta_1 \neq \beta_2 = \beta_3$); and (c) Non-Proportional Odds ($\beta_1 \neq \beta_2 \neq \beta_3$). X is one dimensional and $Y = \{1,2,3,4\}$.

complexity, the main limitations in PO and NPO respectively. The key assumption of the PPO model structure is that a *subset* of predictors satisfies the proportional odds assumption, while the assumption is relaxed for the others:

$$log\left(\frac{\pi_k}{1-\pi_k}\right) = \alpha_k + \sum_{i\in S_1} \beta_i X_i + \sum_{j\in S_2} \gamma_{jk} X_j \qquad (9)$$

$$log\left(\frac{\pi_k}{\pi_{k+1}}\right) = \alpha_k + \sum_{i\in S_1} \beta_i X_i + \sum_{j\in S_2} \gamma_{jk} X_j \qquad (10)$$

$$log\left(\frac{\pi_k}{\sum_{l=k+1}^{K} \pi_l}\right) = \alpha_k + \sum_{i\in S_1} \beta_i X_i + \sum_{j\in S_2} \gamma_{jk} X_j \qquad (11)$$

The equations (9), (10) and (11) are PPO structures for COM, ACM and CRM respectively. S_1 is the set of predictors that uphold the PO assumption, while the predictors in set S_2 violate the assumption. PPO overcomes the burden of assuming a common relationship for all categories across all predictors and hence avoiding possible underfitting, while restricting the number of model parameters relative to NPO (which may overfit). In order to fit a PPO model structure, S_1 and S_2must be chosen prior to modelling.

COM-PPO and even COM-NPO are susceptible to stochastic order violation. Stochastic ordering is defined as follows: for any two given observations x and y, if $G(p(x_i)) < G(p(y_i))\ \forall i$, then x is stochastically higher than y. When stochastic ordering is violated, predicted probability could be negative. In statistics, if the stochastic order is violated, the feature space is limited to the region where the stochastic ordering is preserved. However, ACM and CRM are not bounded by the stochastic constraints and hence always provide valid probabilities for PPO and NPO models.

4 Proposed PPO Model Selection

4.1 Testing the Proportional-Odds Assumption

When fitting either COM, ACM or CRM models, the proportionality structure should be chosen beforehand, i.e. depending on whether the dataset completely holds the assumption, violates the assumption or partially holds the assumption will determine which proportionality structure (PO, NPO or PPO) is best suited. According to the definition of PPO (Section 3.4), it provides a total of $\lambda = \sum_{r=1}^{D-1} C_r^D$ model combinations, where D is the number of predictor variables and r is the number of predictors violating the proportional-odds assumption. With PO and NPO, the total number of proportionality structures offered by logistic regression framework is $2 + \lambda$. Choosing the correct proportionality structure from this vast model space is vital, and incorrect choice of model structure could result in low prediction accuracy. One study [72] has shown how using PO when the proportionality assumption is violated can lead to incorrect conclusions for diabetic patients.

The choice of proportionality structure depends on the proportional-odds assumption: upholding it provides proportional-odds whilst violation of the assumption completely or partially leads to NPO and PPO respectively. To validate the PO assumption, statistical tools such as the likelihood-ratio test, score test and Wald test have been utilised [70]. These tools compare PO against NPO or an augmented model (fitting binary logistic regression for each dichotomous partition pair). Hence these tools validate the assumption in a global context but not on the individual predictor variables. Therefore, they only allow choice between PO and NPO.

Another naive approach to proportionality model structure selection is to compare the coefficients of the augmented models [70, 72]: if the coefficients β for a particular predictor variable are constant across all depression severity levels, it suggests the upholding of the proportional odds assumption, if not violation of the assumption. A drawback of this approach is that there is no clear procedure to validate whether variation of the coefficients is sufficiently large in order to warrant a violation of the assumption.

An augmented model based solution to test the proportionality assumption by incorporating a Wald-type goodness-of-fit statistic was presented in [73], which is referred to as the *Brant test*. Unlike the above mentioned conventional statistical tests, the Brant test assesses the proportionality assumption for individual predictor variables, and hence assists in PPO model selection. However, the original Brant test only accommodated COM. An adaptation of the Brant test for ACM was presented in [74], but CRM has still not been investigated to date. Moreover, although the Brant appears widely in statistical literature, it has not been used with high-dimensional large datasets.

Although statistical tools to test the proportional odds assumption exist, still a general approach for proportionality structure selection, that is applicable to any ordinal logistic regression model and does not resort to an exhaustive search, is missing from the literature.

4.2 GREP: Proposed Greedy Proportionality Structure Selection

GREP is inspired by the concept of hierarchical clustering, where one element is merged with the cluster at a time. Similarly, GREP will decide the proportionality structure by considering one predictor variable at a time using a heuristic generated from the augmented model. It is important to note that the purpose of GREP is not to evaluate the proportional-odds assumption with respect to predictor variables but to find the '*best*' proportionality structure. This is achieved by evaluating the fitness of the model at each step

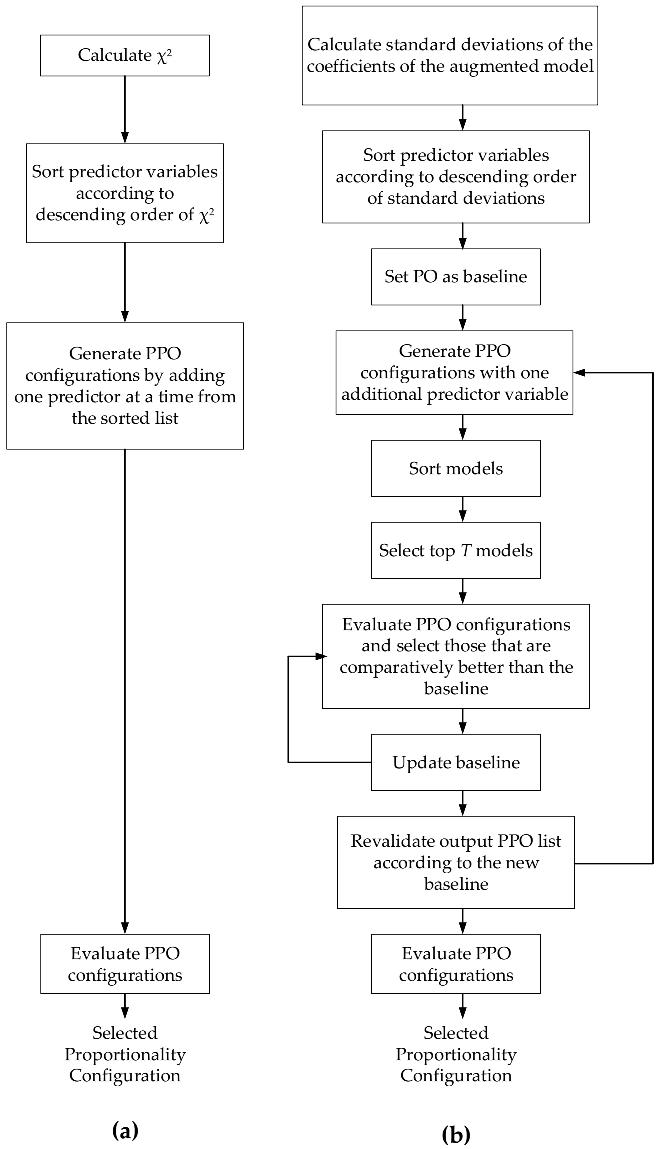
Calculate χ²
Sort predictor variables according to descending order of χ²
Generate PPO configurations by adding one predictor at a time from the sorted list
Evaluate PPO configurations
Selected Proportionality Configuration
(a)
Calculate standard deviations of the coefficients of the augmented model
Sort predictor variables according to descending order of standard deviations
Set PO as baseline
Generate PPO configurations with one additional predictor variable
Sort models
Select top T models
Evaluate PPO configurations and select those that are comparatively better than the baseline
Update baseline
Revalidate output PPO list according to the new baseline
Evaluate PPO configurations
Selected Proportionality Configuration
(b)

Fig. 4: Summary of PPO model selection using (a) the Brant test and (b) GREP. The two algorithms use two different statistics (Brant: χ^2, GREP: standard deviation) to sort predictors. GREP follows a few additional steps to generate a set of PPO configurations whereas in the Brant test, model configurations are generated directly from chi-squared values.

using the Akaike Information Criterion (AIC) (see Fig. 4)

$$\mathrm{AIC} = -2\ln(LL) + 2Q, \qquad (12)$$

where LL is the maximum likelihood of the fitted model and Q is the number of model parameters, which depends on the dimensionality of the feature set, the number of levels of the depression scale and the type of proportionality configuration. The output of the algorithm could be either the proportionality structure with the lowest AIC or a set of model structures whose AICs are not so different. GREP adopts the AIC criterion because AIC considers both model fitness and complexity and hence allows less room for overfitting. Even still, AIC can in principle be replaced with any evaluation measure if it is more suitable for the context of the problem.

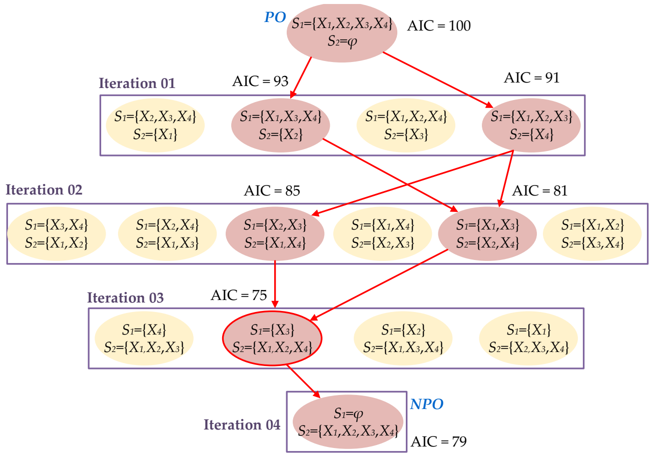

Fig. 5. Illustration of PPO model selection in GREP using an example of a 4-dimensional dataset where only X_3 follows the PO assumption. T = 0.4 and γ_{aic}= 0. In the first iteration, only the second and fourth models (red colour), having a lower AIC than the initial PO model, are considered. Ultimately GREP only evaluates 7 model configurations (red colour) out of the 16 possible model configurations (yellow colour). The algorithm has converged to the correct model structure for the dataset (circled in red), where $S_1 = \{X_3\}$ and $S_2 = \{X_1, X_2, X_4\}$. In the fourth iteration GREP still assesses the NPO structure, however since its AIC is higher than that of the previous model structure, NPO is ignored.

GREP starts with the simplest proportionality structure, PO, so that at the beginning all the predictors are in the set S_1. Then GREP moves one variable at a time to set S_2 (forward step) only if the new model has a lower AIC than the baseline AIC (β_{aic}). At the beginning, β_{aic} is initialised to the AIC of PO model. The flow diagram of GREP is given in Fig. 4(b), whilst the algorithm is presented in Fig. 6.

The model structure to consider in each iteration is chosen in a greedy manner. First, the coefficients of the augmented models are calculated. The standard deviations of the coefficients (std_{φ_d}) are taken as a heuristic for sorting the predictor variables. Large standard deviations suggest violation of the assumption and hence predictors with larger standard deviations are assigned a higher priority. Based on the priority of each variable in the selected model configuration(g), a ranking score (r_g) is calculated. The models with the top $T \in [0,1]$ percentage are selected for the evaluation. The sorting algorithm is presented in Fig. 7.

The margin $\gamma_{aic} \in (-\infty, +\infty)$ is used to control the flexibility at the baseline (β_{aic}). γ_{aic} can be either positive, zero or negative. $\gamma_{aic} = 0$ is the general scenario where only models with AIC less than or equal to the baseline are accepted. $\gamma_{aic} \in \mathbb{R}^{+}$ rejects any improvements smaller than the margin whereas $\gamma_{aic} \in \mathbb{R}^{-}$ accepts degradations smaller than or equal to the margin. If AIC is within the accepted range, the model is added to the output list M. At the end of each iteration, the list M is revalidated to make sure the selected model structures are globally optimal (backward step). The number of models evaluated in each iteration is controlled by hyper-parameters T and γ_{aic}. The application of the GREP algorithm to a synthetic dataset is depicted in Fig. 5.

```
algorithm Greedy Proportionality Structure Selection
input: set of feature vectors $X \in \mathbb{R}^D$, set of labels $Y$, model type $P$
output: selected PPO model structures $M$

$\tau \leftarrow TRUE$
$M \leftarrow \phi$
$L \leftarrow$ sort predictors

$\alpha_{aic} \leftarrow$ calculate AIC for PO
$M \leftarrow M \cup PO$
$\beta_{aic} \leftarrow \alpha_{aic}$

while $\tau = TRUE$ do
    $G \leftarrow$ generate all models with one additional predictor variable
    for each model $g \in G$ do
        $r_g \leftarrow$ summation of ranking indices (decreasing) of predictors

    $G_{sorted} \leftarrow$ sort($G$, $r_g$)
    $C \leftarrow$ select top $T$ models from $G_{sorted}$

    for each PPO/ NPO model $c \in C$ do
        $\alpha_{aic} \leftarrow$ calculate AIC for $c$
        if $\left( \left\{ \frac{\beta_{aic} - \alpha_{aic}}{\beta_{aic}} \right\} \geq \gamma_{aic} \right)$ then
            $M_0 \leftarrow M_0 \cup c$
            if $\alpha_{aic} < \beta_{aic}$ then
                $\beta_{aic} \leftarrow \alpha_{aic}$
        else
            continue

    if $M_0 \equiv \phi$
        $\tau \leftarrow FALSE$
    else
        $M \leftarrow M \cup M_0$
$M \leftarrow$ revalidate $M$ with new $\beta_{aic}$ and remove not good models
```

Fig. 6. GREP: Greedy Proportionality Structure Selection Algorithm

```
algorithm sort predictors
input: set of feature vectors $X \in \mathbb{R}^D$, set of labels $Y$, model type $P$
output: sorted predictors $L$

$\varphi \leftarrow$ calculate coefficients of the augmented model
for each $d \in \{1, \cdots, D\}$ then
  $std_{\varphi_d} \leftarrow$ standard_deviation($\varphi_d$)
$L \leftarrow$ sort($std_{\varphi}$,descending)
```

Fig. 7. Sorting algorithm for predictor variables

5 Experimental Settings

5.1 Databases

In this study two databases were considered: the AViD (Audio-Visual Depressive) corpus, a frequently used depression corpus in the literature, and the DAIC-WOZ (Distress Analysis Interview Corpus - Wizard of Oz) [75] corpus, which is the largest publicly available depression database so far. Both these two depression databases are slightly skewed towards the lower end of the scale and do not have examples for some depression scores. The two databases are summarised in Table 1 with respect to the number of sessions and minimum, maximum and average length of recordings. For further information/ statistics regarding the databases please refer to: AViD [20, 76], DAIC-WOZ [22, 77]. We acknowledge that not having a separate validation partition is a limitation with such small datasets and may produce somewhat optimistic results.

A subset of the AViD corpus, which contains audio and video recordings of interviews with patients suffering from depression, was the database provided for the AVEC 2013 challenge [20]. Interviews were conducted by a computer agent, and the recordings consist of 14 different tasks including vowel phonation, speaking out loud, counting, read speech, singing, telling a story (imagined or past experience) and answering questions. The participants were native German speakers. The AViD corpus is labelled using BDI-II depression scale.

The DAIC-WOZ corpus, a subset of which was employed in the AVEC 2017 challenge [22], is a multimodal corpus containing audio, video and text, but only audio recordings were used in this work. This corpus consists of interviews with patients conducted by a virtual agent. The audio files contain the speech of both interviewer and interviewee, in English. The PHQ-8 depression scale was used to annotate the DAIC-WOZ corpus.

5.2 Experimental Settings

Since DAIC-WOZ audio files contain speech of both interviewer and patient, for the experiments herein, the patients' speech was isolated using the transcripts provided with the database. Ordinal logistic models only accommodate a consecutive range of ordinal scores, but both databases have missing samples for some depression scores. Furthermore, some depression levels have one or few samples only. Therefore, we followed the following groupings to train the models, DAIC-WOZ (PHQ-8): 0-1, 2-4, 5-9, 10-14, 15-19, 20-24, AViD (BDI): 0-1, 2-3, 4-5, 6-7, 8-10, 11-13, 14-20, 21-26, 27-38, 39-63. Without losing the comparability the predicted Y was mapped back to the relevant original depression scale via the mid-point of each grouping prior to model evaluation. Data in both AViD and DAIC-WOZ have already been partitioned into training, development and test. However, we combined the original train and development partitions of AViD in order to get a much bigger training dataset. In DAIC-WOZ, development partition was used as the test partition because ground truth of test partition is not available (Table 1).

The experimental evaluations presented in Section 6 employed four feature sets: MFCC, eGeMAPS low level descriptors (LLDs), SCF and PLLR. MFCC features consisted of 12 coefficients and energy coefficient (C0), appended with delta (Δ) and delta-delta ($\Delta\Delta$) coefficients. MFCCs were extracted from frames of length of 25ms with 10ms shift using the openSMILE toolkit [78]. 15-dimensional SCF features were calculated using a mel-scale Gabor filter bank. For the experiments in this paper, we used eGeMAPS LLDs instead of functionals. Frame-level features were used for these experiments considering the high number of model parameters of both OLR and the baseline

TABLE 1
SUMMARY OF SPEECH DEPRESSION DATABASES USED

DB	Partition	No of Sessions	Min Length (min)	Max Length (min)	Avg. Length (min)
AViD	Train (Train + Dev)	100	7.77	27.33	15.82
	Test	50	5.25	23.01	15.95
DAIC-WOZ	Train	107	6.92	32..77	15.08
	Test (Dev)	35	10.15	27.28	17.16

system, CNN. GSR was similarly originally proposed with frame-level MFCC features. 25-dimensional eGeMAPS LLDs were extracted using openSMILE. PLLR features were generated using pre-trained models using the BUT [79]. For DAIC-WOZ, 39-dimensional English PLLR was computed. Even for AViD, though the recordings are in German, English PLLR was generated because the BUT phoneme recogniser does not support German, followed by stepwise feature selection. A recent study in affective computing found that matching the language for PLLR is not necessary to achieve good performance in continuous emotion prediction [34]. All the audio feature sets were aggregated using a 5s length averaging window with a step size of 1s, for smoothed frame-level features. We applied z-norm to standardise the features for both OLR and CNN.

Ordinal logistic regression models were trained using the R VGAM package [80]. Since the audio feature sets used herein are frame-level, the OLR models predict a depression score for each frame. To derive the final depression score, frame-level predictions were multiplied by the relative frequency of occurrence of that depression score, pre-calculated across the entire dataset. This approach was motivated by the hypothesis that not every frame may equally manifest the depression information and hence higher weights for increasing levels of predicted depression were assigned. Herein, the distribution of depression scores across the dataset was considered an approximation to the distribution of frame-level depression scores for any given utterance.

The baseline GSR model was designed with 5 steps, where the low-class partitions comprised scores of 0-2, 0-5, 0-7, 0-11, 0-15 for the DAIC-WOZ corpus and 0-4, 0-11, 0-18, 0-32, 0-39 [30] for the AViD corpus. The high-class partitions were the complements of the low-class partitions. For comparison purposes, we used GSR without speaker adaptation. The baseline CNN system had a similar architecture (DCNN+DNN) to the system presented in [26], with activation function *ReLU* and *Mean Squared Error* loss function. The system was trained using the *Adam* optimiser. Other hyper-parameters including learning rate, weight decay, batch size were determined empirically. The depression score mapping function followed by OLR models was not required for the baseline systems since both GSR and CNN models are regression models and missing scores can still be predicted.

To assess and compare the model performances, two evaluation measures were exploited in this study. The error-based evaluation measure, RMSE (Root Mean Squared Error), is to-date the most common evaluation measure in the depression prediction literature. Since the depression scores are ordinal, ranking measures are more attractive than error-based evaluation measures [81]. Hence the second evaluation measure adapted was the Spearman Rank Correlation Coefficient (ρ). GREP was tuned on the test partition, unless otherwise stated. All the results are reported on test partition.

6 RESULTS AND DISCUSSION

Section 3 presented an analytical comparison of the three ordinal regression models: COM, ACM and CRM. An empirical analysis of various OLR models (Section 6.1-6.2) as well as a comparison of model selection algorithms (Section 6.3) using two real datasets: DAIC-WOZ and AViD is described in this section. Section 6.4 is devoted to evaluating performance of the proposed algorithm GREP. Finally, a comparison of OLR models with two state-of-the-art systems for depression prediction: GSR and CNN is provided in Section 6.5.

6.1 Comparison of Different Ordinal Logistic Regression Models using Depression Corpora

This section compares the three OLR models: COM, ACM and CRM in terms of PO and NPO only. Each of the OLR models has its own desirable characteristics (Section 3.1): ACM is preferred when the ordinal categories are substantive (section 3.1.2), whereas CRM is favoured when the ordinal scale is sequential (section 3.1.3). According to these guidelines and considering the fact that both AViD and DAIC-WOZ corpora consist of recordings from multiple different speakers, the most suitable model type a priori is ACM.

In terms of PO proportionality structures, eGeMAPS and SCF with ACM can be preferred for the AViD corpus whereas MFCC and SCF with COM perform better on the DAIC-WOZ corpus. For PLLR and eGeMAPS on DAIC-WOZ and MFCC and PLLR on AViD, CRM reported the lowest RMSE. Surprisingly for NPO proportionality structure, all feature sets from both corpora except PLLR on AViD favour CRM. According to the observations on OLR types for AViD and DAIC-WOZ corpora, CRM is the preferred model type to predict depression severity of group of people, since it is the most common model choice for majority of PO and NPO proportionality structures.

It is of particular note that, in terms of RMSE and Spearman correlation, the behaviour of the three types of ordinal logistic regression models are not radically different. NPO configuration model fitting for COM failed with most feature sets and hence COM NPO has been omitted from Fig. 8. COM NPO is highly susceptible to stochastic order violation, and possibly could lead to failures in model fitting.

6.2 Comparison of Proportionality Configurations using Depression Corpora

PPO is favourable when compared with PO and NPO because of the possibility to compromise between model parsimony (PO is the most parsimonious) and the level of model detail (NPO has the highest model complexity). Even still, the exact proportionality structure should be chosen according to the given data. Fig 8. also presents a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3031300, IEEE Transactions on Affective Computing

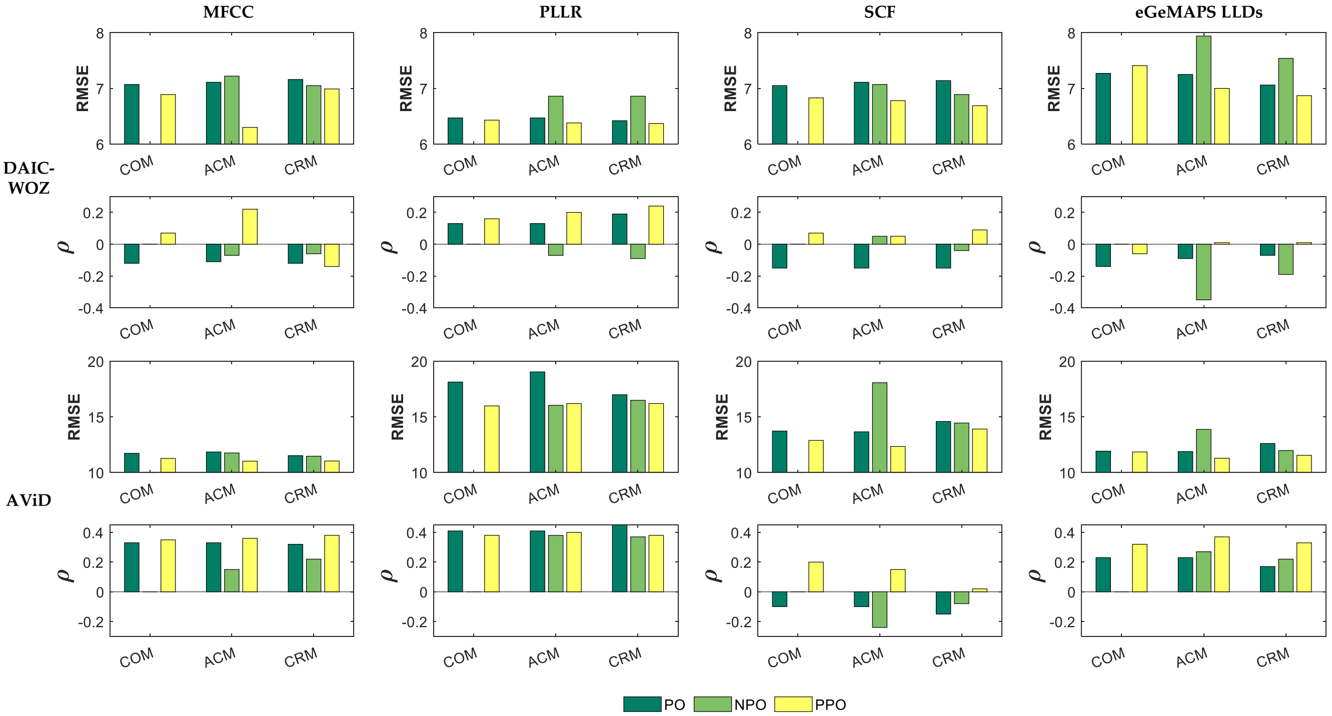

Fig. 8. Depression prediction accuracy of ordinal logistic regression models for three model types: Cumulative-Odds Model (COM), Adjacent Category Model (ACM), Continuation-Ratio Model (CRM) and three proportionality structures: Proportional-Odds (PO), Non- Proportional Odds (NPO), Partial Proportional Odds (PPO). The four columns represent the four different feature sets: MFCC, PLLR, SCF and eGeMAPS LLDs. The top and bottom two rows present RMSE and Spearman correlation evaluated on DAIC-WOZ and AViD.

comparison between different proportionality structures for the AViD and DAIC-WOZ corpora. PPO proportionality structures shown in Fig. 8 were selected using the proposed GREP algorithm.

According to Fig 8, majority of features and corpora, the lowest RMSE and the highest Spearman correlation were recorded with a PPO proportionality structure. Rather than accepting the proportional odds assumption (PO) or completely ignoring the assumption (NPO), relaxing the assumption on some of the predictors (i.e. PPO) resulted in improved performance (reduced RMSE and increased Spearman correlation) for both corpora.

The Spearman correlations for PO and NPO were mostly negative, meaning that ground truth and predicted depression scores has an opposite relationship. As a result of the heterogeneous nature of depression and subjectivity inherent in depression assessment, automatic depression prediction is still an evolving research field and hence negative correlation is observed. Negative correlation of this kind has been reported in a previous study on another challenging affective computing problem [82].

6.3 PPO Model Selection

This section presents a comparison between exhaustive model selection, the Brant test and the GREP algorithm. To begin with, we evaluated the RMSEs of all combinations of partial proportional odds models (exhaustive) and used them as a baseline to compare both the Brant test and the proposed algorithm.

For the purposes of comparison, the first six MFCCs 1-6 coefficients (due to high computational burden) were used throughout the following experiments, with evaluation on the DAIC-WOZ database.

6.3.1 *Exhaustive Model Selection*

According to Fig. 9, the behaviours of COM, ACM and CRM are quite similar when continuously relaxing the proportional-odds assumption. At each stage, there are multiple PPO configurations that give lower RMSE than both PO and NPO, even with CRM which has the lowest RMSE for PO and NPO. The lowest RMSEs for PPO model configurations were as follows: COM 6.22 ($S_2 = \{C1, C4, C5, C6\}$), ACM 6.25 ($S_2 = \{C1, C4, C6\}$), and CRM 6.25 ($S_2 = \{C1, C5, C6\}$).

This type of experiment is not feasible for model selection in general, because it involves evaluating $\lambda + 2$ different model combinations, which scales extremely poorly for high-dimensional features. In this example, the dimension of the feature set was 6 and hence the total number of configurations including PO and NPO was 64, which is feasible to run. However, if the complete MFCC feature set was used, with dimension 39, then the total number of model configurations is approximately 5×10^{11} and hence will take years to exhaustively compute all configurations.

Therefore, it is essential to devise an algorithm that can find a model with good accuracy within a reasonable running time. Since Fig. 9 shows that there are PPO configurations that are much better and much worse than PO and NPO configurations, the algorithm ideally should converge to a model that produces approximately the lowest RMSE.

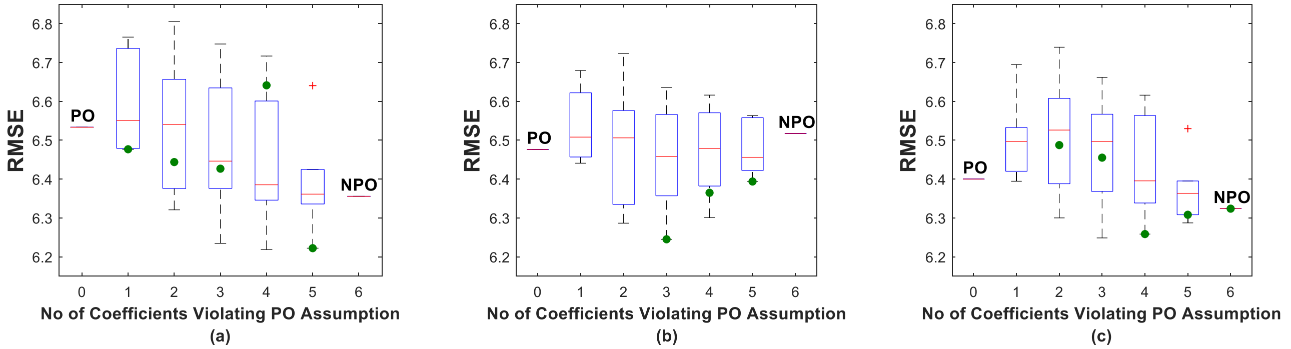

Fig. 9. Boxplots (red and blue) of RMSEs summarising all possible proportional odds model structures of (a) Cumulative-Odds (b) Adjacent-Category and (c) Continuation-Ratio models (Section 6.3.1). PPO has the potential to provide much more accurate proportionality structures than PO or NPO in terms of RMSE. The green dots represent the RMSEs for the output proportionality structures selected using GREP (Section 6.3.3)

6.3.2 Model Selection using Brant Test

This section evaluates the Brant test for the purpose of PPO model selection. According to the Brant test results on MFCC 1-6 on DAIC-WOZ corpus, all the coefficients reject the null hypothesis (results are not reported due to space limitation), i.e. the assumption of parallelism is violated. However, relaxing all predictors (NPO) doesn't give the best RMSE. Even still, the chi-squre statistic could still be useful in determining a PPO which has lower RMSE than PO and NPO.

For this particular example, RMSE has a convex-like relationship (see Fig. 10) when sorted in descending order of chi-squre statistic. However, it is not always the case; RMSE could take more complex relationships with multiple points of inflection, such that finding the proportionality structure with the lowest RMSE is more challenging. Since the total number of models proposed by the Brant test is $(D+1)$, where D is the dimension of the feature set, in this study we evaluated all PPO proportionality structures proposed by the Brant, to determine the *'best'* configuration.

The lowest RMSE observed with the Brant test is 6.25, when $s_2 = \{C1, C4, C6\}$. This is not the lowest RMSE for the dataset, which is 6.22 (Section 6.3.1). However, the observed lowest RMSE is lower than both PO (6.53) and NPO (6.33). The model search space of this approach is considerably narrow than the total model space which is $\lambda + 2$. Therefore, by exapnding the model space it will be possible to find a better-performing proportionality structure.

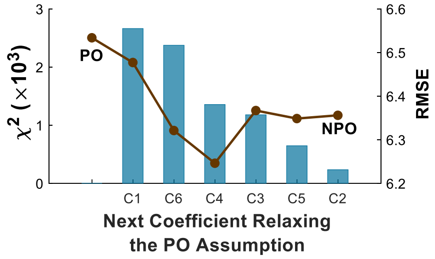

Fig. 10. Brant test summary (COM) for MFCC 1-6 on DAIC-WOZ corpus: Bars indicates chi-square test statistics and brown line depicts RMSE variation, when relaxing the proportional odds assumption for one variable at a time. Left-most and right-most models correspond to PO and NPO respectively. The convex relationship between χ^2 statistic and RMSE suggests that χ^2 statistic could be helpful in finding a PPO structure which is better performing than both PO and NPO.

6.3.3 Model Selection with GREP

GREP results for MFCCs 1-6 are shown in Fig. 9 (green). For this experiment, we used $T = \{0.4, 0.5\}$ and $\gamma_{aic} = \{-0.01, -0.02\}$. Since GREP outputs a set of possible proportionality configurations, we chose the model with the lowest RMSE on training partition per each category: relaxing the proportional-odds assumption with the same number of parameters. GREP has been able to locate a proportionality structure from a lower RMSE range, except when relaxing 4 predictors in COM. This anomalous selection could be due to overfitting, since even though RMSE on training partition is low, RMSE on the test is much higher.

The lowest RMSE PPO configurations observed with GREP were as follows: COM 6.22($S_2 = \{C1, C2, C4, C5, C6\}$), ACM 6.25 ($S_2 = \{C1, C4, C6\}$) and CRM 6.26 ($S_2 = \{C1, C4, C5, C6\}$). GREP has therefore outperformed the Brant test: GREP has been able to find the lowest RMSE PPO configuration for both COM and ACM. Even for CRM, the configuration suggested by GREP is the second best PPO configuration.

6.4 Detailed Evaluation of GREP

The GREP algorithm has two critical hyper-parameters: threshold value T and margin γ_{aic}. With these two hyper-parameters, GREP allows expanding/shrinking the model search space as needed, making GREP more attractive than exhaustive search and the Brant test. The following experiments were designed to evaluate the impact of its hyper-parameters to the proposed algorithm. We used the same set of features as Section 6.3 and evaluated on DAIC-WOZ. When GREP outputs a set of possible models, the lowest RMSE model on training partition was selected.

6.4.1 Effect of Margin (γ_{aic}) on Model Performance

This experiment was conducted to evaluate the effect of margin (γ_{aic}) on the performance of GREP at multiple T values. According to Fig. 11, for smaller T (e.g. T=0.1), γ_{aic} doesn't have any major impact on RMSE. When T is

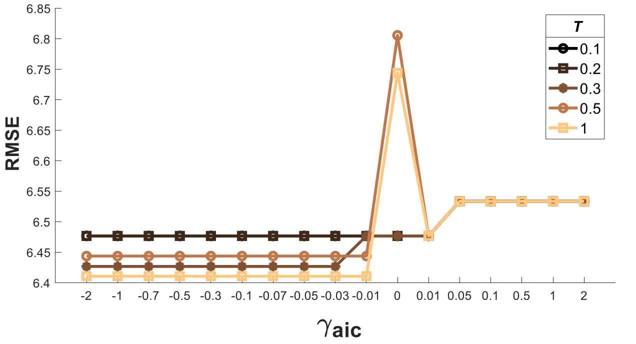

Fig. 11. Variation of RMSE with γ_{aic} values at different T values. When γ_{aic} is slightly negative, lower RMSE can be observed than positive γ_{aic}.

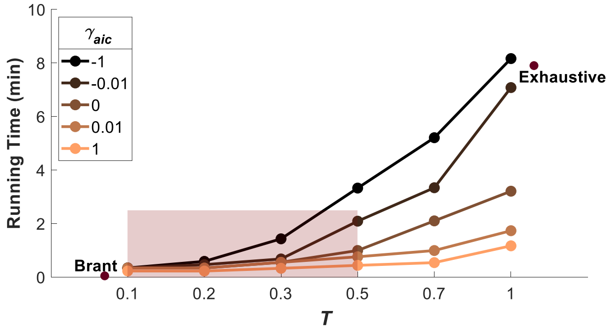

Fig. 12. Running time of GREP algorithm at different T and γ_{aic} on DAIC-WOZ for a 6-dimensional feature set. The left-most and right-most coordinates correspond to running time of the Brant test and exhaustive searches respectively. The range of T used in our experiments is highlighted in red colour.

sufficiently large, choosing a negative γ_{aic} helps in finding low-RMSE PPO proportionality structures.

Negative γ_{aic} allows models with higher AIC than baseline AIC. AIC is one of the most frequently used goodness-of-fit criteria for model evaluation in statistics. Yet, AIC is only capable of evaluating one aspect of the model error [81] and a better system in terms of RMSE might be found with a higher AIC than that of the baseline. On the other hand, according to Fig. 11, choosing a negative γ_{aic} with larger magnitude is not necessary and will consume more time since the model space is unnecessarily expanded. Hence the value of γ_{aic} needs to be chosen with care. For all other experiments reported in this paper, $\gamma_{aic} \in [-0.1, -0.01]$ was used.

6.4.2 Scaling of Running Time with Threshold Value:T

Figure 12 shows the variation in runtime when increasing T at different γ_{aic} along with the running time of the Brant test and exhaustive searches. The running time when T is small ($T < 0.1$) is closer to that of the Brant test. On the other hand, when $T = 1$ and $\gamma_{aic} \rightarrow -\infty$, GREP is equivalent to exhaustive search.

The hyper-parameter T allows flexibility in trading off running time with accuracy. When GREP is equivalent to exhaustive search ($T = 1$ and γ_{aic} is sufficiently large), it will evaluate all possible model configurations and predict the globally optimum model configuration but with the cost of much longer running time. Setting $T < 1$ does not always guarantee the globally optimum model configuration, but a sub-optimal model configuration which is possibly better than PO is predicted. For example in Fig. 11 when $T = 0.3$ and $\gamma_{aic} = -0.03$, RMSE is 6.222 which is not much different from the lowest RMSE (6.22) which is observed with two hyper-parameter settings 1) $T = 0.5$ and $\gamma_{aic} = -0.03$ and 2) $T = 1$ and $\gamma_{aic} = -0.03$. Setting T too small is not recommended as it could yield poor accuracy (not being able to find a model with lower RMSE than PO). For the experiments reported in this paper, $T \in [0.1, 0.3]$ was used.

6.5 Model Comparison with Baseline Systems

Finally, we compared the ordinal logistic regression models with partial proportional odds (PPO) structures selected by the proposed GREP algorithm with two state-of-the-art systems: GSR and CNN (Table 2). GSR shares some conceptual similarities with COM (Section 2), and hence is a good baseline for the OLR framework. DNN-based systems have gained attention in the depression prediction community due to their high accuracy [19, 26]. Hence, we also included a CNN model as a baseline system, specifically the best-performing system [26] from the AVEC2017 challenge.

For the DAIC-WOZ database, the performance of OLR was better than that of GSR and CNN in terms of both RMSE and Spearman correlation except for eGeMAPS LLDs. For the AViD corpus, GSR reported the lowest RMSE for both SCF and eGeMAPS LLDs whereas CNN outperformed other two systems with PLLR. However, in terms of Spearman correlation OLR outperformed GSR and CNN for MFCC and eGeMAPS LLDs. In Table 2, the two symbols * and + indicate the statistical significance of Spearman correlation coefficient vs. the null hypothesis (no correlation) at two levels: (p<0.05) and (p<0.1) respectively. There is a discrepancy between RMSE and Spearman correlation results in Table 2. These are two conceptually different types of evaluation measures which evaluate the error from two different perspectives [81]. Spearman correlation which is a ranking measure, is apt for assessing ordinal systems [81]. Therefore, from Table 2, it is fair to conclude that with careful choice of hyper-parameter settings ordinal logistic regression can outperform GSR and CNN.

Performance differences between these three systems were statistically validated using one-tailed Wilcoxon signed ranks tests. A series of 30 trials were conducted per feature set, randomly dividing into train and test partitions in each trial. Two separate tests were conducted to compare OLR against GSR and CNN for both RMSE and Spearman correlation per feature set. According to Wilcoxon tests ($p < 0.1$), OLR is significantly better than GSR and CNN in terms of Spearman correlation for DAIC-WOZ. Both GSR and CNN are state-of-the-art systems, hence results for which significance was not found nevertheless represent strong performance from OLR.

7 Conclusion

This paper has investigated the logistic regression framework for the purpose of designing an ordinal regression

TABLE 2
COMPARISON OF DEPRESSION PREDICTION ACCURACY (RMSE AND SPEARMAN'S CORRELATION) OF ORDINAL LOGISTIC REGRESSION WITH PROPOSED GREP MODEL SELECTION AND BASELINE SYSTEMS (GSR AND CNN)

	DAIC-WOZ							
	MFCC		PLLR		SCF		eGeMAPS LLDs	
	RMSE	RHO	RMSE	RHO	RMSE	RHO	RMSE	RHO
GSR	7.37	-0.36*	6.75	**0.45***	7.24	-0.26^{+}	6.99	-0.10
CNN	6.78	-0.06	6.74	-0.12	6.73	0.01	**6.84**	**0.11**
OLR	**6.30**	**0.22**$^{+}$	**6.37**	0.24^{+}	**6.69**	**0.09**	6.87	0.01
	AViD							
	MFCC		PLLR		SCF		eGeMAPS LLDs	
	RMSE	RHO	RMSE	RHO	RMSE	RHO	RMSE	RHO
GSR	11.57	**0.36***	11.57	0.28^{+}	**11.58**	0.30^{+}	**11.28**	0.34*
CNN	11.84	0.05	**11.37**	**0.44***	11.77	**0.36***	11.34	0.32*
OLR	**11.03**	**0.36***	15.99	0.38	12.34	0.15	11.29	**0.37***

*Significance of Spearman's rho is indicated as follows: * represents p-value < 0.05 and + represents p-value < 0.1*

model for depression prediction. So far in the affective computing literature, logistic regression has been used only for 2-class classification. Most affective computing problems including depression, anxiety, emotions and cognitive load, are fundamentally ordinal, due to conventional measures being based on human subjective opinion, and there has been growing community interest in ordinal machine learning approaches. Even still, the application of ordinal logistic regression for affective computing has not yet been investigated despite it being one of the few standard, well-established ordinal regression models available. Moreover, existing studies on ordinal logistic regression (in other machine learning and statistic domains) have been limited to small datasets and rarely been extended to large, more realistic datasets such as the DAIC-WOZ or AViD depression corpora.

Ordinal logistic regression has three model types: COM, ACM and CRM and three proportionality structures: PO, NPO and PPO. Each of the three model types follows different logit formation strategies and hence captures different properties of the ordinal scale. PPO is the intermediate structure between PO and NPO, which gives the most design flexibility, allowing a trade-off between model complexity and parsimony. As seen in Fig. 8, it also has the potential to be the most accurate proportionality structure. PPO contributes to more than one model configurations (In Fig. 9, except left and rightmost models which are PO and NPO respectively, all the others are PPO configurations) and hence introduces the necessity of a judicious model selection algorithm. In this study, we have proposed such a model selection algorithm, and compared it with two existing model selection algorithms.

An exhaustive model selection approach always provides the globally optimal model configuration, but it scales catastrophically with the dimension of the feature set. The Brant also has limitations: it is not generalised for all OLR model types and has a limited model search space. GREP provides an effective approach to make model selection tractable for real problems and is a new contribution to both affective computing and to ordinal logistic regression in general. Results on DAIC-WOZ and AViD attest that GREP can provide a better proportionality configuration than PO or NPO (Section 6.3.3).

With PPO and GREP, ordinal logistic regression can achieve RMSE and Spearman correlation results that are comparable with or improve on state-of-the-art depression prediction systems, GSR and CNN. This is remarkable because OLR is a much simpler model than any deep neural network architecture. Moreover, ordinal regression is the most suitable approach to solve ordinal problems rather than regression. Therefore, ordinal logistic regression should be an essential framework for many affective computing applications.

ACKNOWLEDGMENT

Julien Epps was supported by ARC Linkage grant Australian Research Council Linkage Project LP160101360, and Julien Epps and Eliathamby Ambikairajah were supported by Data61, CSIRO.

REFERENCES

[1] W. H. Organization, "Depression and other common mental disorders: global health estimates," 2017.

[2] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A Review of Depression and Suicide Risk Assessment using Speech Analysis," *Speech Communication*, vol. 71, pp. 10-49, 2015.

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

[4] D. O'shaughnessy, *Speech Communication: Human and Machine*. Universities Press, 1987.

[5] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling spectral variability for the classification of depressed speech," in *Interspeech*, 2013, pp. 857-861.

[6] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and PTSD," in *Interspeech*, 2013, pp. 847-851.

[7] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on

motor incoordination," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013: ACM, pp. 41-48.

[8] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller, "CCA based feature selection with application to continuous depression recognition from acoustic speech features," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014: IEEE, pp. 3729-3733.

[9] A. T. Beck, R. A. Steer, and M. G. Carbin, "Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation," *Clinical psychology review,* vol. 8, no. 1, pp. 77-100, 1988.

[10] K. Kroenke, R. L. Spitzer, J. B. Williams, and B. Löwe, "The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review," *General hospital psychiatry,* vol. 32, no. 4, pp. 345-359, 2010.

[11] 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.

[12] P. McCullagh, "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 109-142, 1980.

[13] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014: ACM, pp. 65-72.

[14] N. Cummins, J. Epps, V. Sethu, and J. Krajewski, "Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014: IEEE, pp. 970-974.

[15] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering,* vol. 58, no. 3, pp. 574-586, 2011.

[16] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, "Probabilistic acoustic volume analysis for speech affected by depression," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[17] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision tree based depression classification from audio video and language information," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016: ACM, pp. 89-96.

[18] A. Pampouchidou *et al.*, "Depression assessment by fusing high and low level features from audio, video, and text," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016: ACM, pp. 27-34.

[19] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An Efficient Deep Model for Audio Based Depression Classification," in *Proc. of the 6th Intl. Workshop on Audio/Visual Emotion Challenge*, 2016: ACM, pp. 35-42.

[20] M. Valstar *et al.*, "AVEC 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013: ACM, pp. 3-10.

[21] V. Mitra *et al.*, "The SRI AVEC-2014 evaluation system," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014: ACM, pp. 93-101.

[22] 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: ACM, pp. 3-9.

[23] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, "Relevance vector machine for depression prediction," in *Sixteenth annual conference of the international speech communication association*, 2015.

[24] H. Pérez Espinosa, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y-Gómez, D. Pinto-Avedaño, and V. Reyez-Meza, "Fusing Affective Dimensions and Audio-Visual Features from Segmented Video for Depression Recognition: INAOE-BUAP's Participation at AVEC'14 Challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014: ACM, pp. 49-55.

[25] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble cca for continuous emotion prediction," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014: ACM, pp. 19-26.

[26] 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: ACM, pp. 53-59.

[27] F. Ringeval *et al.*, "AVEC 2019 Workshop and Challenge: State-of-Mind, Depression with AI, and Cross-Cultural Affect Recognition," in *Proc. of the 9th Annual Workshop on Audio/Visual Emotion Challenge*, 2019.

[28] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors," *Speech Communication,* vol. 52, no. 1, pp. 12-40, 2010.

[29] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[30] N. Cummins, J. Epps, V. Sethu, and J. Krajewski, "Weighted Pairwise Gaussian Likelihood Regression for Depression Score Prediction," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015: IEEE, pp. 4779-4783.

[31] S. D. Østergaard, S. Jensen, and P. Bech, "The heterogeneity of the depressive syndrome: when numbers get serious," *Acta Psychiatrica Scandinavica,* vol. 124, no. 6, pp. 495-496, 2011.

[32] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing,* vol. 7, no. 2, pp. 190-202, 2016.

[33] 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: IEEE, pp. 960-964.

[34] Z. Huang and J. Epps, "An Investigation of Partition-based and Phonetically-aware Acoustic Features for Continuous Emotion Prediction from Speech," *IEEE Transactions on Affective Computing*, 2018.

[35] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "On the use of phone log-likelihood ratios as features in spoken language recognition," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012: IEEE, pp. 274-279.

[36] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3031300, IEEE Transactions on Affective Computing

and G. Bordel, "Using phone log-likelihood ratios as features for speaker recognition," *evaluation,* vol. 3, p. 15, 2013.

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

[38] 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.

[39] J. Fürnkranz and E. Hüllermeier, "Pairwise preference learning and ranking," in *European conference on machine learning*, 2003: Springer, pp. 145-156.

[40] K. Crammer and Y. Singer, "Pranking with ranking," in *Advances in neural information processing systems*, 2002, pp. 641-647.

[41] R. Herbrich, T. Graepel, and K. Obermayer, *Regression Models for Ordinal Data: A Machine Learning Approach*. Citeseer, 1999.

[42] W. Chu and Z. Ghahramani, "Preference learning with Gaussian processes," in *Proceedings of the 22nd international conference on Machine learning*, 2005: ACM, pp. 137-144.

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

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

[45] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of machine learning research,* vol. 4, no. Nov, pp. 933-969, 2003.

[46] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker, "Label ranking by learning pairwise preferences," *Artificial Intelligence,* vol. 172, no. 16-17, pp. 1897-1916, 2008.

[47] H. Zhang, L. Jiang, and J. Su, "Augmenting naive bayes for ranking," in *Proceedings of the 22nd international conference on Machine learning*, 2005: ACM, pp. 1020-1027.

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

[49] 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.

[50] F. Fernández-Navarro, A. Riccardi, and S. Carloni, "Ordinal neural networks without iterative tuning," *IEEE transactions on neural networks and learning systems,* vol. 25, no. 11, pp. 2075-2085, 2014.

[51] W.-Y. Deng, Q.-H. Zheng, S. Lian, L. Chen, and X. Wang, "Ordinal extreme learning machine," *Neurocomputing,* vol. 74, no. 1-3, pp. 447-456, 2010.

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

[53] J. C. Huhn and E. Hüllermeier, "Is an ordinal class structure useful in classifier learning?," *IJDMMM,* vol. 1, no. 1, pp. 45-67, 2008.

[54] S. Jayawardena, J. Epps, and E. Ambikairajah, "Support Vector Ordinal Regression for Depression Severity Prediction," in *International Conference on Information and Automation for Sustainability (ICIAFS)*, Sri Lanka, 2018: IEEE.

[55] N. Brummer *et al.*, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 7, pp. 2072-2084, 2007.

[56] S. Pigeon, P. Druyts, and P. Verlinde, "Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions," *Digital Signal Processing,* vol. 10, no. 1-3, pp. 237-248, 2000.

[57] Z. Huang *et al.*, "Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016," in *Proc. of the 6th Intl. Workshop on Audio/Visual Emotion Challenge*, 2016: ACM, pp. 19-26.

[58] J. F. Cohn *et al.*, "Detecting depression from facial actions and vocal prosody," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 2009: IEEE, pp. 1-7.

[59] H. Jiang *et al.*, "Detecting Depression Using an Ensemble Logistic Regression Model Based on Multiple Speech Features," *Computational and Mathematical Methods in Medicine,* vol. 2018, 2018, doi: 10.1155/2018/6508319.

[60] J.-T. Chien and S. Furui, "Predictive hidden Markov model selection for speech recognition," *IEEE Transactions on Speech and Audio Processing,* vol. 13, no. 3, pp. 377-387, 2005.

[61] D. Ververidis and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm," in *2005 IEEE International Conference on Multimedia and Expo*, 2005: IEEE, pp. 1500-1503.

[62] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, 1998, vol. 2: IEEE, pp. 645-648.

[63] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association,* vol. 96, no. 454, pp. 746-774, 2001.

[64] M. H. Hansen and B. Yu, "Minimum description length model selection criteria for generalized linear models," *Lecture Notes-Monograph Series,* pp. 145-163, 2003.

[65] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2009.

[66] J. J. Verbeek, N. Vlassis, and B. Kröse, "Efficient greedy learning of Gaussian mixture models," *Neural computation,* vol. 15, no. 2, pp. 469-485, 2003.

[67] S. Axelrod and B. Maison, "Combination of hidden Markov models with dynamic time warping for speech recognition," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. 1: IEEE, pp. I-173.

[68] T. P. Minka, "Algorithms for maximum-likelihood logistic regression," 2012.

[69] A. S. Fullerton and J. Xu, "Constrained and Unconstrained Partial Adjacent Category Logit Models for Ordinal Response Variables," *Sociological Methods & Research,* vol. 47, no. 2, pp. 169-206, 2015, doi: 10.1177/0049124115613781.

[70] A. Agresti, *Analysis of Ordinal Categorical Data*. John Wiley & Sons, 2010.

[71] B. Peterson and F. E. Harrell Jr, "Partial proportional odds models for ordinal response variables," *Applied statistics,* pp. 205-217, 1990.

[72] R. Bender and U. Grouven, "Using binary logistic regression models for ordinal data with non-proportional odds," *Journal of clinical epidemiology,* vol. 51, no. 10, pp. 809-816, 1998.

[73] R. Brant, "Assessing Proportionality in the Proportional Odds

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3031300, IEEE Transactions on Affective Computing

Model for Ordinal Logistic Regression," *Biometrics,* pp. 1171-1178, 1990.

[74] A. Dolgun and O. Saracbasi, "Assessing proportionality assumption in the adjacent category logistic regression model," *Statistics and its Interface,* vol. 7, no. 2, pp. 275-295, 2014.

[75] J. Gratch *et al.*, "The Distress Analysis Interview Corpus of human and computer interviews," in *LREC*, 2014: Citeseer, pp. 3123-3128.

[76] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: a multimodal approach," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 11-20.

[77] L. Zhang, J. Driscol, X. Chen, and R. Hosseini Ghomi, "Evaluating Acoustic and Linguistic Features of Detecting Depression Sub-Challenge Dataset," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 47-53.

[78] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013: ACM, pp. 835-838.

[79] P. Schwarz, P. Matejka, L. Burget, and O. Glembek. "Phoneme Recognizer based on Long Temporal Context." Faculty of Information Technolog, Brno University of Technology. https://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context (accessed.

[80] T. W. Yee, "The VGAM package for categorical data analysis," *Journal of Statistical Software,* vol. 32, no. 10, pp. 1-34, 2010.

[81] S. Jayawardena, J. Epps, and E. Ambikairajah, "Evaluation Measure for Depression Prediction and Affective Computing," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* 2019: IEEE.

[82] 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.

Sadari Jayawardena received the BSc (Eng) degree in Computer Science and Engineering from University of Moratuwa, Moratuwa, Sri Lanka in 2014. Currently she is pursuing her Ph.D. degree with Signal Processing Group in the School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, Australia. Her current research interests include machine learning and speech-based depression prediction. She is a member of IEEE, ISCA, IEEE Signal Processing Society and Engineers Sri Lanka.

Julien Epps (M'97) received the B.E. and Ph.D. degrees from the University of New South Wales, Sydney, NSW, Australia, in 1997 and 2001, respectively. From 2002 to 2004, he was a Senior Research Engineer with Motorola Labs, where he was engaged on speech recognition. From 2004 to 2006, he was a Senior Researcher with National ICT Australia, Sydney. He then joined the School of Electrical Engineering and Telecommunications at the University of New South Wales, Australia, as a Senior Lecturer in 2007, and is now Professor and Head of School. He is also a Contributed Researcher at Data61, CSIRO, Australia. He has authored or co-authored more than 250 publications and serves as an Associate Editor for IEEE Transactions on Affective Computing. His current research interests include characterization, modeling, and classification of mental state from behavioral signals, such as speech, eye activity, and head movement.

Professor Eliathamby Ambikairajah received his BSc (Eng) (Hons) degree from the University of Sri Lanka, and received his PhD degree in Signal Processing from Keele University, UK. He was appointed as Head of Electronic Engineering and later Dean of Engineering at the Athlone Institute of Technology in the Republic of Ireland from 1982 to 1999. His key publications led to his repeated appointment as a short-term Invited Research Fellow with the British Telecom Laboratories, U.K., for ten years from 1989 to 1999. Professor Ambikairajah is currently serving as the Acting Deputy Vice-Chancellor Enterprise, after previously serving as the Head of School of Electrical Engineering and Telecommunications, University of New South Wales (UNSW), Australia from 2009 to 2019. As a leader he has firmly established the School as the top Electrical Engineering school in Australia and among the top 50 in the world. His research interests include speaker and language recognition, emotion detection and biomedical signal processing. He has authored and co-authored approximately 300 journal and conference papers and is the recipient of many competitive research grants. For his contributions to speaker recognition research, he was a Faculty Associate with the Institute of Infocomm Research (A*STAR), Singapore in 2009-2018, and is currently an Advisory Board member of the AI Speech Lab at AI Singapore. Professor Ambikairajah was an Associate Editor for the IEEE Transactions on Education from 2012-2019. He received the UNSW Vice-Chancellor's Award for Teaching Excellence in 2004 for his innovative use of educational technology and innovation in electrical engineering teaching programs, and again in 2014 he received the UNSW Excellence in Senior Leadership Award and in 2019 he was the recipient of the People's Choice Award as part of the UNSW President's Awards. Professor Ambikairajah was an APSIPA Distinguished Lecturer for the 2013-14 term. He is a Fellow and a Chartered Engineer of the IET UK and Engineers Australia (EA) and is a Senior Member of the IEEE and a Life Member of APSIPA.