Psychological Assessment

Detecting Depression Using a Framework Combining Deep Multimodal Neural Networks With a Purpose-Built Automated Evaluation

Ezekiel Victor, Zahra M. Aghajan, Amy R. Sewart, and Ray Christian

Online First Publication, May 2, 2019. http://dx.doi.org/10.1037/pas0000724

American Psychological Association

Psychological Assessment

1040-3590/19/$12.00

http://dx.doi.org/10.1037/pas0000724

Detecting Depression Using a Framework Combining Deep Multimodal Neural Networks With a Purpose-Built Automated Evaluation

Ezekiel Victor
Textsavvyapp, Inc., West Hollywood, California

Zahra M. Aghajan
Textsavvyapp, Inc., West Hollywood, California, and University of California, Los Angeles

Amy R. Sewart
University of California, Los Angeles

Ray Christian
Textsavvyapp, Inc., West Hollywood, California

Machine learning (ML) has been introduced into the medical field as a means to provide diagnostic tools capable of enhancing accuracy and precision while minimizing laborious tasks that require human intervention. There is mounting evidence that the technology fueled by ML has the potential to detect and substantially improve treatment of complex mental disorders such as depression. We developed a framework capable of detecting depression with minimal human intervention: artificial intelligence mental evaluation (AiME). This framework consists of a short human-computer interactive evaluation that utilizes artificial intelligence, namely deep learning, and can predict whether the participant is depressed or not with satisfactory performance. Because of its ease of use, this technology can offer a viable tool for mental health professionals to identify symptoms of depression, thus enabling a faster preventative intervention. Furthermore, it may alleviate the challenge of observing and interpreting highly nuanced physiological and behavioral biomarkers of depression by providing a more objective evaluation.

Public Significance Statement
The current study presents a novel paradigm that uses machine learning (multimodal deep networks) to detect depression. This framework can function as a screening tool for depression and be integrated into clinical settings to assist mental health professionals.

Keywords: depression, artificial intelligence, deep learning, mental health evaluation, multimodal classification

Machine learning (ML), a method of data analysis in which computers learn to independently modify or adapt their actions (e.g., make predictions) to produce more accurate decisions and results, has emerged as a powerful analytic tool for large and complex data sets (Marsland, 2011). As such, ML lends itself to the processing of disease biomarkers and has been implemented in medical diagnostic tools ranging from the detection and classification of tumors (Petricoin & Liotta, 2004; Bocchi, Coppini, Nori, & Valli, 2004) to providing a differential diagnosis of neurodegenerative diseases with similar presentations (Salvatore et al., 2014). ML methods have reliably demonstrated an increase in prediction accuracy when compared with older, more conventional statistical techniques or physician-based expert systems (Cruz & Wishart, 2007).

In parallel, ML has been applied to examine affective display differences exhibited during emotion states, such as facial expression and vocal prosody, through audio and video-based analyses. These advances have generated a field of research that has successfully used ML techniques, such as support vector machines (Cohn et al., 2009), regression (Valstar et al., 2013), and neural networks (L. Yang et al., 2017), for automatic emotion recognition using audiovisual data (Schuller, Steidl, & Batliner, 2009;

Ezekiel Victor, Textsavvyapp, Inc., West Hollywood, California; Zahra M. Aghajan, Textsavvyapp, Inc., and Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles; Amy R. Sewart, Department of Psychology, University of California, Los Angeles; Ray Christian, Textsavvyapp, Inc.

We thank the participants for volunteering to participate in our study.

This work was supported by Textsavvyapp, Inc., and much of the data processing was done with the help of Azure server credits provided by Microsoft through their generous BizSpark Plus program.

Ezekiel Victor and Ray Christian are full-time employees of Textsavvyapp, Inc. Zahra M. Aghajan is a part-time employee of Textsavvvyapp, Inc.

Correspondence concerning this article should be addressed to Zahra M. Aghajan, Textsavvyapp, Inc., 907 North Harper Avenue, Suite 8, West Hollywood, CA 90046. E-mail: zahra@textpert.ai

Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005; Dhall, Goecke, Joshi, Wagner, & Gedeon, 2013; McKeown, Valstar, Cowie, Pantic, & Schroder, 2012; Ringeval, Sonderegger, Sauer, & Lalanne, 2013). Moreover, ML has been extended to investigate verbal and nonverbal affective abnormalities associated with psychiatric disorders and has shown promise in detecting measurable differences between those presenting with and without a given diagnosis (Hamm, Kohler, Gur, & Verma, 2011; P. Wang et al., 2008; Gratch et al., 2014; Alhanai, Ghassemi, & Glass, 2018). This is a substantial advancement, given that prior to the advent of ML, identifying divergences in affect-related behaviors relied exclusively on labor-intensive, rater-based analysis (Gaebel & Wölwer, 1992), which can be susceptible to raters' personal biases.

ML-based techniques show promise for psychiatric diagnostics by harnessing observable affect-related behaviors through objective methods. In fact, observable affect-related behaviors are commonly used by mental health professionals to assist in psychiatric diagnostics, often through unstructured methods that result in general, qualitative data (e.g., flat or broad affect). ML algorithms' reliance on quantitative and observable behaviors is therefore compelling and applicable for clinical use. Unfortunately, the majority of current algorithms still require some level of human intervention such as labor-intensive manual labeling or hand classification of data to extract useful features prior to analysis (Valstar et al., 2013; Valstar et al., 2014).

We sought to investigate the possibility of developing a method that combines advanced ML-based techniques in combination with automated data collection procedures to identify clinical depression in a demographically diverse population. We chose to begin this effort with depression for two reasons. First, the prevalence and impact of depression is staggering and affects millions of people across the globe. Depression is the leading cause of disability in the United States for individuals ranging from 15 to 44.3 years of age (Substance Abuse and Mental Health Services Administration, 2017). Major depressive disorder, a psychiatric disorder characterized by experiencing depressed mood or anhedonia most of the day nearly every day for a period of 2 weeks or more, affects upward of 16.2 million American adults annually, roughly 6.7% of the United States population (Substance Abuse and Mental Health Services Administration, 2017). Distress from clinically elevated depression is often accompanied with suicidal ideation and attempt (World Health Organization, 2017). Nearly 800,000 individuals worldwide die as the result of suicide each year, making it the second leading cause of death in individuals 15–29 years of age. Second, verbal and nonverbal affective abnormalities demonstrated by individuals with depression are well documented and lend themselves to ML processing. Depressed individuals possess significant differences in facial expressions (Girard & Cohn, 2015) and everyday vocabulary use (e.g., absolutist words; Al-Mosaiwi and Johnstone, 2018) when compared with healthy individuals. In addition, speaking behaviors and voice acoustic characteristics (e.g., F_0 and switching pauses; Yang, Fairbairn, & Cohn, 2013) have been closely linked to depressive state, recovery time course from depression (Kuny & Stassen, 1993), and treatment response (Mundt, Snyder, Cannizzaro, Chappie, & Geralts, 2007). This research provides a solid foundation of behavioral biomarkers that may be used to identify clinically elevated depression using audiovisual data.

Hence, we designed a web-based evaluation that can be completed quickly (~5 min), requires no manual labeling, and takes into account all of the above-mentioned modalities. In addition, we created a new ML-based algorithm that leverages and extends the behaviorally relevant findings to identify depression using naturalistic audiovisual data. This comprehensive methodology (artificial intelligence mental evaluation, AiME) was developed to minimize human intervention, thereby enhancing feasibility, scalability, and potential applications in clinical settings.

Method

Participants

We collected data from 671 participants who performed a human-computer interactive evaluation. Participants were recruited through Amazon Mechanical Turk and were compensated for their time. The evaluation (completed at participants' homes or their preferred location) was primarily composed of interview questions where participants were recorded by a webcam and a microphone while they responded to questions relating to their mental well-being. The evaluation also contained an anonymous demographics questionnaire (age, sex, ethnicity, etc.) as well as a brief, multiple-choice, mental health questionnaire to provide additional data and ground-truth validation. Participants were asked to confirm that recording requirements (lighting, camera angle, etc.) were met (see Appendix).

The evaluation took approximately 5 minutes, and data from the demographics questionnaire, video responses, and mental health questionnaires were stored and accessed in accordance with Health Insurance Portability and Accountability Act compliance standards. All participants were from the United States. We performed automated validations to ensure adequate data quality (e.g., each video contained a single clearly recognizable face and voice). Videos of particularly poor visual and/or audio quality were discarded as judged case by case. The final reported participant numbers are after the cleanse (i.e., final data included for training and testing). The resulting sample of participants was 57.97% female, 41.73% male, and 0.30% other; 73.77% White, 10.13% African American, 8.35% Hispanic/Latino, 4.47% Asian/Pacific Islander, 0.59% Native American, 0.45% Middle Eastern, and 2.24% other; 3.43% in the age range of 15–19 years, 23.99% in 20–24 years, 22.95% in 25–29 years, 33.38% in 30–39 years, 9.69% in 40–49 years, 4.62% in 50–59 years, and 1.94% in 60 years and above.

This study did not seek institutional review board approval because the data in the current study were collected as part of product development by Textsavvyapp, Inc. This type of research without academic affiliation has precedence in industry; nonetheless, the product has been subsequently adopted as a part of an institutional review board–approved clinical trial.

Measures

Video questions. Participants responded vocally to eight questions regarding current mental well-being for 15–60 s per question (e.g., "How have you been feeling lately?"). Similarly, participants responded vocally to five additional questions regarding past and current treatment history for 3–30 s per question (e.g.,

"Has a mental health professional diagnosed you with depression in the past?"). During these questions, video and audio data were collected. Participant's behavioral data from the first eight questions recorded via video and audio as well as speech content (what was said) were used for prediction. The full list of these interview questions is included in Appendix.

Depression. Participants completed the Patient Health Questionnaire (PHQ-9) (Kroenke, Spitzer, & Williams, 2001), a nine-item self-report measure that assesses depression on a 4-point scale (from 0 = *not at all* to 3 = *nearly every day*). Total scores range from 0 to 27, with higher scores denoting a greater endorsement of depressive symptoms. Scores from PHQ-9 were used as the "ground truth" for the training and assessment of models.

Statistical Approach

We developed a multimodal deep learning model that used video data, audio data, and speech content from participants' responses as well as demographics and other metadata. These data were used as adjacent inputs to the model to predict depression in two ways: (a) treating depression as a continuous problem and thus using regression (predicting PHQ-9 scores between 0 and 27 against the true PHQ-9 values); and (b) performing binary classification on whether participants were likely to be considered clinically depressed (using a cutoff value on the PHQ-9 scores that is used in frequent clinical practice).

Data processing involved the following steps: (1) Video data were subsampled to eight frames per second, cropped to participants' face (using Google Cloud Vision) and then downsampled to 128 $\times$ 128 pixels (Figure 1A) and finally analyzed using an architecture resembling ResNet (He, Zhang, Ren, & Sun, 2016); (b) audio data from the microphone were downsampled to 80 Hz, and 22 features were extracted over the entire time trace. These features included 13 Mel-frequency cepstral representations as well as other features such as spectral roll-off, entropy, and so forth (Figure 1A); and (c) speech content was automatically transcribed using Google Cloud Speech service and transformed to word representation vectors using Global Vectors (GloVe 6B) (Pennington, Socher, & Manning, 2014)—and subsequently used as another input to the model (Figure 1A). These data streams were passed through long short-term memory (LSTM) recurrent neural network (Hochreiter & Schmidhuber, 1997) layers because of the time-varying nature of the inputs. Lastly, the model combined these inputs with a dense layer containing demographic information and other metadata (for each video, the metadata vector consisted of the total duration of the video, word count, unique word count, and word density defined as the number of words per second), and prediction occurred after the application of dense layers (Figure 1B).

We applied data augmentation, a commonly used method that simultaneously increases the number of input data points as well as reduces the potential for overfitting the model (i.e., making the outcome invariant to geometric and color properties of individual images; Wang & Perez, 2017). In particular—in addition to providing raw video to the model—we mirrored the video (geometric) and adjusted color contrast (color). Lastly, the scores from the PHQ-9 were used as the ground truth. Computations were implemented using Keras (TensorFlow backend). We report our findings from one regression model (that also allowed us to compare our results with those from prior work in the literature) as well as two binary classification models for which a PHQ-9 score of 10 was used as a threshold for depression.

The regression model was trained on 537 exams using a mean squared error loss function and an independent set of 134 exams was left for testing (true PHQ-9 scores for 671 exams: 7.95 $\pm$ 6.48, mean $\pm$ *SD*; for 537 exams in training set = 7.98 $\pm$ 6.50; for 134 exams in test set = 7.89 $\pm$ 6.42). Similarly, the classification models were trained on 365 exams (balanced; i.e., equal number of depressed vs. nondepressed exams) using a binary cross-entropy loss function and an independent set of 91 exams was left for testing (for the test set, the base rate for the binary classification variable was 33.58%, i.e., the percentages of exams that exceeded the PHQ-9 threshold; true PHQ-9 scores: 4.02 $\pm$ 2.82 for PHQ-9 $<$ 10 and 15.53 $\pm$ 4.26 for PHQ-9 $\geq$10 classes, mean $\pm$ *SD*).

It is important that, for classification problems, training occurs on a balanced data set, that is, equal number of examples in each class, to prevent training a naive model that always predicts the majority class in practice. Thus, to balance our training data set, we used data from all depressed participants and an equal number of exams from the nondepressed category (this resulted in discarding 215 exams). The test set, on the other hand, was not balanced and was sampled from the original distribution of exams to ensure the sample is a true representation of the population distribution. In both cases, the 80/20 split for the training and testing data was done randomly and to ensure there were enough data in each group to minimize variances in both parameter estimates and performance metrics. The group assignments were done based on participant, and all model manipulations—training the algorithm, model tuning, and optimization—were done solely on the training data (and blind to the test set that was left only for testing performance metrics). As an additional precaution, random group assignment was performed before each isolated experiment (as opposed to fixed assignment over the duration of the study) to ensure that model tuning was agnostic to the particular composition of the evaluated test set. In addition to the holdout method described, we performed 10-fold cross-validation (with validation and test sets) to eliminate the possibility of overfitting to a single holdout test set and obtain more accurate performance estimates (see *Results*).

As a further precaution against overfitting to the training data, we used early stopping (Yao, Rosasco, & Caponnetto, 2007), a regularization method that limits the number of training iterations. In the regression model, the output of the model (predicted y) was compared against the true PHQ-9 scores. In the classification model, the output of the model (predicted y values were between 0 and 1) was rounded to construct a binary vector consisting of ones (depressed) and zeros (nondepressed) and was compared against the true binarized PHQ-9 scores. The second classification model was different from the first in that we also performed hyperparameter optimization (hyperparameters are human-predefined parameters as opposed to parameters learned by the model during training) using random search (Bergstra & Bengio, 2012) and we used bidirectional LSTM (BiLSTM) in lieu of LSTM.

Results

We implemented a regression model in which the model outputs were trained against the PHQ-9 scores (0–27); we used a scaled

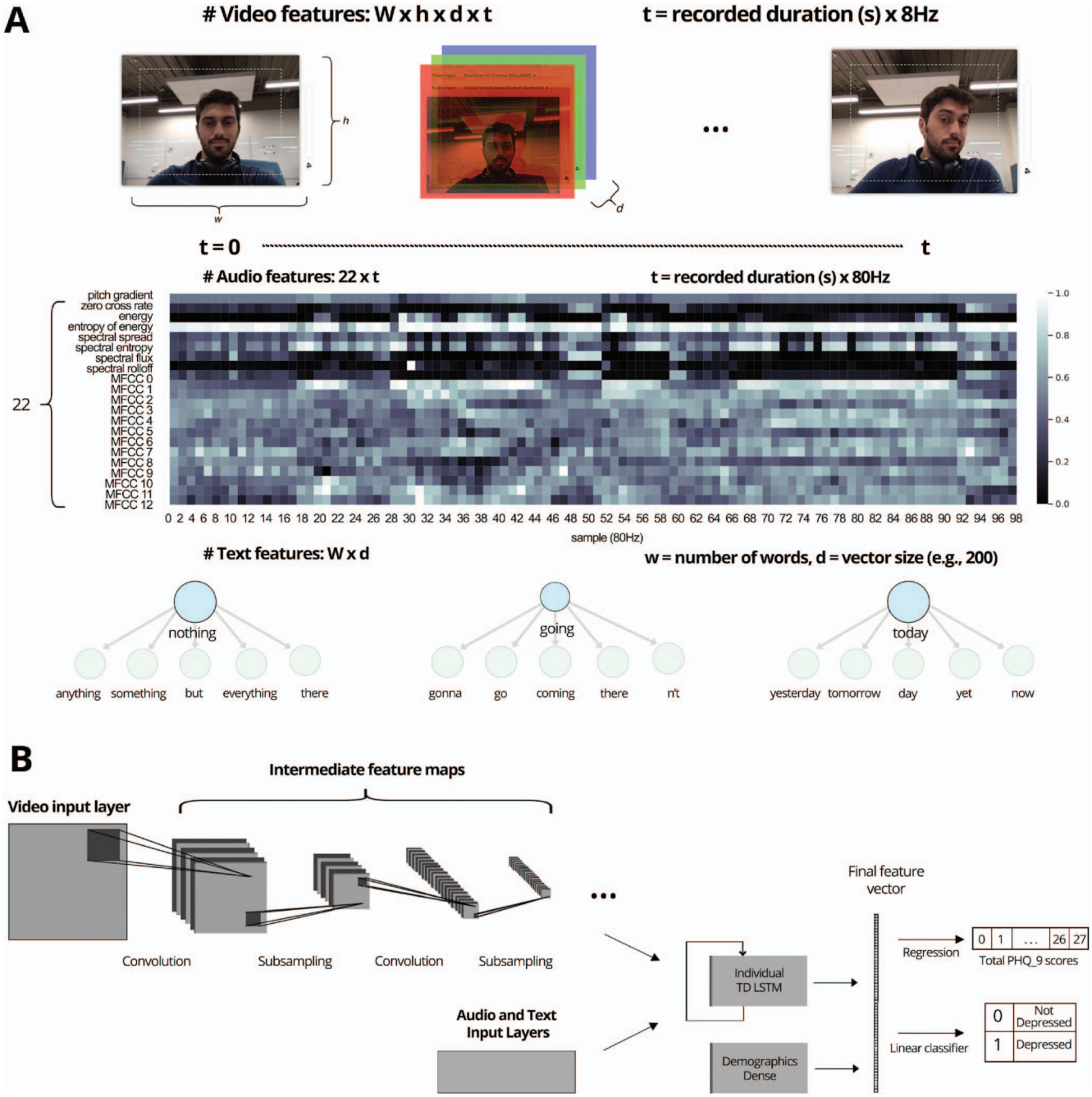

Figure 1. Inputs and the architecture of the network used for classification. Example inputs from three different modalities (video, audio, text). Top, Video data consisted of frames whose dimensions were in 128 × 128 pixels in width and height, and three in depth (corresponding to the RGB color channels). Depending on the total length of the video, the number of video frames per sample also varied (eight frames per second). Middle, Audio data, similar to video data, consisted of frames (80 per second) that were transformed into 22 features (see text), including short-term power representations (mel-frequency cepstrum). Bottom, Text data were analyzed using GloVe representation with a word vector size of 200 (see text), thus leading to input dimensions of number of words × 200 (copyright 2019 by Textsavvyapp, Inc. Adapted with permission) (A). An overview of the network demonstrating how different data streams are processed individually and combined (B). See the online article for the color version of this figure.

(by 27) sigmoid activation function, and a mean squared error loss function (i.e., the model would minimize the mean squared error between the true and predicted PHQ-9 scores). Mean absolute error (MAE) and root mean square error (RMSE) were used as measures for model evaluation: (a) $MAE = \frac{1}{n}\sum_{j=1}^{n} |y_{true} - y_{predicted}|$, (b) $RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} (y_{true} - y_{predicted})^2}$, which also allowed for comparisons with previously described models (see Table 1). Although the use of different scales for training (Yang et al., 2017 used PHQ-8 scores) can introduce confounds for performance comparisons, we scaled our model's MAE and RMSE values to obtain error percentages. We found our model fared better compared to previous models (Yang et al., 2017), although the comparison is indirect because of the PHQ-8 versus PHQ-9 difference. Direct comparisons of our model performance on other data sets (Cohn et al., 2017) warrants further investigation but is beyond the scope of this article.

Curiously we found that when regression models were trained on male and female participants separately they performed remarkably better. This might be due to the fact that the underlying distribution of the PHQ-9 scores was significantly different between female (7, [3, 13]; median, [25th, 75th]%) and male (2, [5, 10]; median [25th, 75th]%) participants ($p = 0.00024$; Wilcoxon rank-sum test). This difference in model performance between male and female participants had also been previously reported (Yang et al., 2017). We also noted that the performance metrics during testing were slightly better than training metrics. A possible explanation for this is that we used the dropout regularization method during training to prevent overfitting. These dropout layers are later deactivated during testing to allow the model full access to intermediate features and may result in better performance during testing.

It is worth noting that we treated audio-visual response data from individual questions (of the eight total questions within the evaluation) as independent observations. However, the ultimate goal is to aggregate these question-level estimates into a single estimate of the patient. To achieve this, we performed two different analyses: (a) we took the average model predictions from all eight questions and computed the model performance by comparing that score against the true PHQ-9 value (MAE test: 4.83 ♀, 4.82 ♂; RMSE test: 6.05 ♀, 6.10 ♂); and (b) we built a simple neural network with two dense layers that used the predictions from eight questions as the input and performed regression to predict the true single PHQ-9 score. Consistent with our prior framework, we had independent sets of training and testing data (MAE test: 4.59 ♀, 4.37 ♂; RMSE test: 5.90 ♀, 5.52 ♂). This analysis is critical, given that although behavioral manifestations of depression may change from moment to moment, the underlying level of depression may not.

Initially model selection was determined by the epoch that resulted in the best performance on the test set. Although reasonable at first blush, this introduces the possibility of choosing a model overfit to the test data. To alleviate this, we later performed

Table 1
Summary of Various Measures Used for the Evaluation of Model Performance

Model 1, Performance metrics (regression error), all data					
Train MAE	Train RMSE		Test MAE	Test RMSE	
5.27 (19.52%)	6.68 (24.74%)		5.12 (18.96%)	6.35 (23.52%)	
Regression model trained and tested on female participants					
4.03 (14.93%)	5.25 (19.44%)		3.73 (13.81%)	4.91 (18.19%)	
Regression model trained and tested on male participants					
3.46 (12.81%)	4.80 (17.78%)		3.55 (13.15%)	4.52 (16.74%)	
Model 2, Performance metrics (classification), balanced model ($\tau = 0.41$)					
Accuracy	(PPV) precision	NPV	Sensitivity	Specificity	F-1 score
68.02, 95% CI [67.32, 68.75] **70.88**	68.61, 95% CI [66.36, 70.81] **80.46**	67.61, 95% CI [66.32, 68.28] **73.46**	68.59, 95% CI [64.57, 72.87] **86.81**	67.46, 95% CI [62.59, 72.60] **87.77**	67.66, 95% CI [65.77, 68.50] **71.15**
90% specific model $\tau = 0.63$					
70.16	82.93		49.93	**89.95**	62.33
90% sensitive model $\tau = 0.28$					
61.40	56.96		**89.79**	33.63	69.70

Note. PPV = positive predictive value; NPV = negative predictive value. For the first model measures, the numbers inside parentheses represent the percentage of error with respect to the range of possible values (27 for PHQ-9 scores); thus, lower numbers indicate better performance. Percentage error is provided to aid in comparison to prior studies such as Yang et al., 2017, which used a PHQ-8 scale having a range of possible values of 24 points, thereby otherwise confounding direct comparison of MAE and RMSE. Values from the second model are reported as bootstrapping results ($n_{iterations} = 10{,}000$) representing the estimated value and 95% confidence intervals ($n_{epochs} = 25$) and were computed during the test phase (i.e., unseen data). Underlined values in bold indicate the highest values obtained from different model epochs. Parameter τ represents the threshold at which the predictions were considered positive. Model 1 (regression; all data) training: $N_{participants} = 537$, $N_{examples} = 4{,}296$; testing: $N_{participants} = 134$; $N_{examples} = 1{,}072$; Model 1 (regression; female) training: $N_{participants} = 312$, $N_{examples} = 2496$; Model 1 testing: $N_{participants} = 77$; $N_{examples} = 616$; Model 1 (regression; male) training: $N_{participants} = 224$, $N_{examples} = 1{,}792$; Model 1 testing: $N_{participants} = 55$; $N_{examples} = 440$; Model 2 training: $N_{participants} = 365$, $N_{examples} = 2{,}920$; Model 2 testing: $N_{participants} = 91$, $N_{examples} = 728$.

10-fold cross-validation with both validation and test sets. Here model selection was guided by performance on the validation set with final model performance quantified by the MAE on the independent test set: (a) validation MAE (4.72, 95% confidence interval [CI; 4.50, 4.94]), test MAE (4.80, 95% CI [4.25, 5.35]), female; and (b) validation MAE (4.82, 95% CI [4.70, 4.95]), test MAE (4.88, 95% CI [4.48, 5.25]), male. This procedure is not only useful for preventing overfitting to the test set during model selection but also results in more accurate estimates of model performance.

Although depression is a continuous phenomenon, it is frequent clinical practice to use cutoffs to dichotomize this continuous measure. One way to achieve this is to apply thresholds on the output of our regression models. Alternatively, we used a classification model that was trained against binarized labels (whether the participant was depressed or not based on applying a threshold on the PHQ-9 scores; for a detailed description see the *Statistical Approach* section). We used various metrics (Fabian et al., 2011) to evaluate the model performance at each epoch (each epoch representing a single full presentation of the data set to the model during training). First, we quantified the receiver-operating characteristics curve and the area under that curve—common measures of model performance in classification (Huang & Ling, 2005)—for each model as well as individual epochs within a model (see Figure 2). The model converged and its performance remained stable for many epochs during which it successfully classified depression labels well above chance (see Figure 2). The final model selection is commonly left to the experimenter to choose the epoch with the highest desired performance metric.

Moreover, we constructed the contingency table (confusion matrix) of the number of true positives (TP; correctly identified as depressed by the model), true negatives (TN; correctly identified as nondepressed by the model), false positives (FP; nondepressed

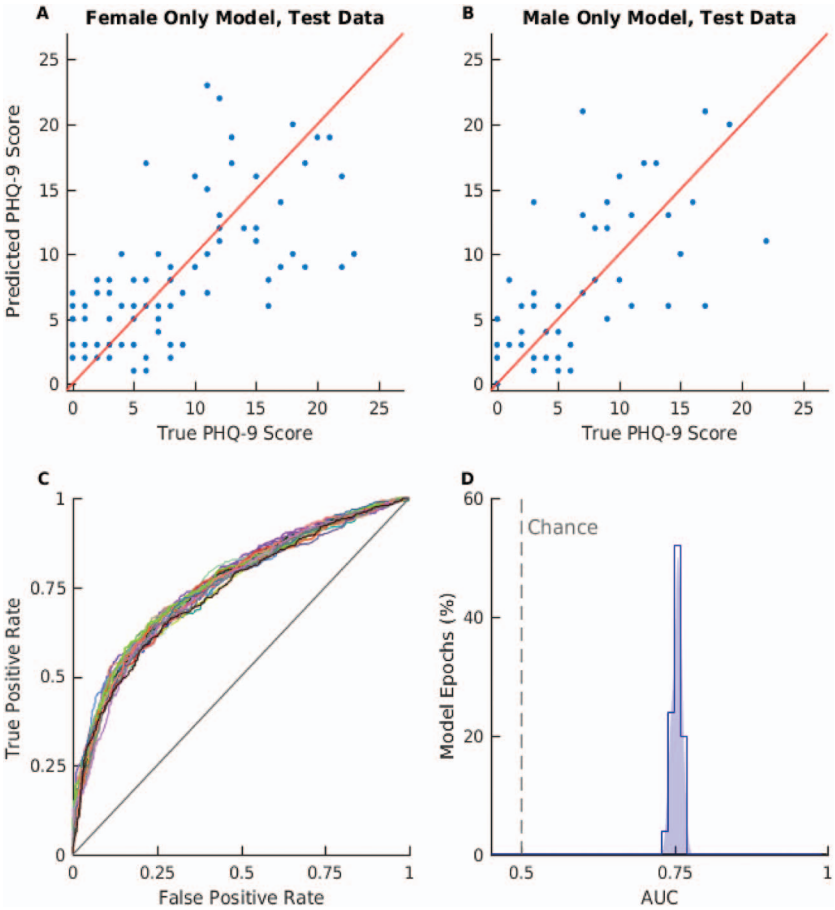

Figure 2. Classification achieved performances well above chance level. Scatter plots are the regression model predictions for the PHQ-9 scores versus the true PHQ-9 scores in the test data. Individual data points are the prediction results combined over 8 questions per participant within each gender group in the test data from a representative epoch (A and B). Receiver operating characteristics (ROC) curve obtained from 25 representative test epochs from a model (each line represents a different epoch) indicating that classification was done above chance level (gray diagonal line) (C). The distribution of the area under the ROC curve (AUC) values (mean ± $SD = 0.75 \pm 0.01$) for the curves shown in C (D). Gray dashed vertical line indicates chance level. These metrics are derived from predictions made during the test phase (i.e., unseen data). See the online article for the color version of this figure.

individuals that were identified as depressed by the model), and false negatives (FN; depressed individuals that were identified as nondepressed by the model). These values were used to compute the following metrics to gauge the model performance (for a detailed description, see Table 1):

(a) accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$, (b) positive predictive value; precision): $\frac{TP}{TP+FP}$, (c) negative predictive value: $\frac{TN}{TN+FN}$, (d) sensitivity (recall): $\frac{TP}{TP+FN}$, (e) specificity: $\frac{TN}{TN+FP}$, and (f) F-1 score: $\frac{2}{\frac{1}{precision}+\frac{1}{sensitivity}}$.

According to the utilized metrics, the described models exhibited satisfactory (above chance) performance levels. The second classification model (random search) outperformed the first model (data not shown) and achieved stable performance across all measures (see Table 1). Two representative epochs reached high specificity and sensitivity values (87.77% and 86.81%, respectively), and in fact, it is possible to adjust the threshold value (τ) at which a prediction is considered positive to achieve higher levels of specificity or sensitivity (see Table 1). It has been argued that it is important to report diagnostic test results at different thresholds, particularly for binary classification problems, because the clinical relevance and optimal threshold values will likely depend on the type of diagnostics performed (Mallett, Halligan, Thompson, Collins, & Altman, 2012).

Discussion

In this article we have introduced a novel methodology, which combines a brief evaluation and ML techniques to detect depression. Our model takes advantage of the fact that there are significant differences in facial expressions, tone of voice, and vocabulary used by individuals with depression compared to the nondepressed population. Our results suggest it is possible to detect depression (or a depressive state) with methods that require minimal human intervention both in terms of data collection and labeling. It must be noted that despite having achieved satisfactory performance levels, there are some limitations that will be detailed below.

One limitation of the current approach is that because the self-report exam is conducted at specific moments in time, the behavioral results might be the individual's state-dependent affect (a short-term emotional influence caused by a recent event), rather than the long-term affective characteristics associated with depression. The exam is brief and accessible, however, and thus can be taken multiple times—for example at periodic time intervals—which can mitigate the state-dependent affect. Completing the evaluation at periodic time intervals may also offer two additional benefits: a longitudinal assessment of the depressive state, and insights into subtle depressive symptom changes over time.

Another limitation—a common caveat of supervised learning methods—is that to train the model to perform a regression or classification task, data need to be labeled (e.g., using a continuous measure of depression or depressed vs. not depressed classes). In our algorithm, the ground truth is determined based on the self-reported, and therefore subjective, PHQ-9 scores that are not fully accurate in depression prediction. Nonetheless, it is possible to expand upon the current method, and utilize labels provided by psychiatrists and mental health specialists, for example from Structured Clinical Interview for D *Diagnostic and Statistical Manual of Mental Disorders*-5 interviews, to obtain potentially better and more accurate performance measures. Moreover, by training the model against a scale like PHQ-9, which is shown to correlate with depression risk and severity (Kroenke et al., 2001), the model is transitively learning to associate time-dependent behavioral biomarkers with depression risk and severity. Furthermore, prediction loss against PHQ-9 is reduced and smoothed by averaging across random batches of predictions prior to model parameter updates in such a way that the model will not practically consider the PHQ-9 scale as unadulterated truth but rather a rough gauge of depression severity. Lastly, as with any method that lacks a direct physician-patient interaction, the models presented here (like PHQ-9) can be seen only as screening or triage tools and cannot yet offer diagnoses by themselves.

Neural networks are often considered black box methods because of obscure inner workings with features and parameters not explicitly guided by human hands. The models described herein are no different, although new research is showing indications of increasing interpretability of results by using methods such as activation map visualization (Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015; Zintgraf, Cohen, Adel, & Welling, 2017; Chen et al., 2018) to understand what influences network decisions.

Another limitation of the study is that the data were not collected in a clinical setting. In some respects this is deleterious, for example because of a lack of personal guidance of study participants by laboratory technicians or physicians, and an inability to ensure maximum recording quality and consistency. In other respects it is useful, for example to train a robust network invariant to implementation details like camera, microphone, lighting quality, and so forth. Another related concern, which is beyond the scope of this article, is that a clinical population may exhibit meaningfully different characteristics compared to a nonclinical population, even at similar PHQ-9 levels.

Taken together, the current study presents a proof of concept for detecting depression severity and risk using ML techniques on behavioral data. Although trained from PHQ-9 scores, ML models have the potential to improve and eventually provide more refined performance levels that go above and beyond current questionnaire-based methods. This is particularly notable because behavioral analysis is potentially less susceptible to report bias often found in self-reported data. Additionally, output from the penultimate model layer, an n-dimensional vector resulting from preceding dense layers of n units, may be interpreted as a qualitative descriptor of the subject input in n-space, and therefore analyzed and compared with clustering techniques (e.g., k-means or t-SNE). The resultant clusters may inform future diagnostic criteria and improve our understanding of depression and myriad other ailments detectable from behavioral biomarkers beyond the capabilities of a traditional questionnaire in clinical settings.

Neurotechnologies (e.g., chronic brain implants) have emerged as viable options that use electrophysiological biomarkers to treat psychiatric diseases such as depression. Artificial intelligence mental evaluation could be used along with these neuromodulation techniques—such as the application of deep brain stimulation and transcranial magnetic stimulation to treat pharmaco-resistant depression (Bewernick et al., 2010; George et al., 2000)—to provide finer precision progress tracking for recipients of novel treatment as well as eventually correlate behavioral changes with brain activity already being gathered in the course of such treatments.

The combination of precision and objectivity with machine learning methodology may offer a chance at enhancing our ability to track and triage individuals suffering depression. Additionally, the ability to automate and thus scale depression diagnostic data collection may support mental health professionals with easier access to robust patient mental health insights.

References

Alhanai, T., Ghassemi, M., & Glass, J. (2018). Detecting depression with audio/text sequence modeling of interviews. *Interspeech* (pp. 1716–1720). Retrieved from http://dx.doi.org/10.21437/interspeech.2018-2522

Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science, 6,* 529–542. http://dx.doi.org/10.1177/2167702617747074

Bergstra, J., & Bengio, Y. (2012). Random *Search for Hyper-Parameter Optimizatio*n. *Journal of Machine Learning Research, 13,* 281–305.

Bewernick, B. H., Hurlemann, R., Matusch, A., Kayser, S., Grubert, C., Hadrysiewicz, B., . . . Schlaepfer, T. E. (2010). Nucleus accumbens deep brain stimulation decreases ratings of depression and anxiety in treatment-resistant depression. *Biological Psychiatry, 67,* 110–116. http://dx.doi.org/10.1016/j.biopsych.2009.09.013

Bocchi, L., Coppini, G., Nori, J., & Valli, G. (2004). Detection of single and clustered microcalcifications in mammograms using fractals models and neural networks. *Medical Engineering & Physics, 26,* 303–312. http://dx.doi.org/10.1016/j.medengphy.2003.11.009

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech)* (pp. 1517–1520). Lisbon, Portugal.

Chen, X., Guan, Q., Lo, L.-T., Su, S., Ahrens, J., & Estrada, T. (2018). In situ TensorView: In situ visualization of convolutional neural networks. ArXiv:1806.07382 [Cs. CV].

Cohn, J., Cummins, N., Epps, J., Goecke, R., Joshi, J., & Scherer, S. (2017). Multimodal assessment of depression and related disorders based on behavioural signals. *Handbook of Multimodal-Multisensor Interfaces, 2,* 375–417.

Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., . . . De la Torre, F. (2009). Detecting depression from facial actions and vocal prosody. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (pp. 1–7). Amsterdam, the Netherlands. http://dx.doi.org/10.1109/ACII.2009.5349358

Cruz, J. A., & Wishart, D. S. (2007). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics, 2,* 59–77.

Dhall, A., Goecke, R., Joshi, J., Wagner, M., & Gedeon, T. (2013). Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 509–516). Sydney, Australia. Retrieved from https://dl.acm.org/citation.cfm?doid=2522848.2531749

Fabian, P., Gaël, V., Alexandre, G., Vincent, M., Bertrand, T., Olivier, G., . . . Alexandre, P. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830.

Gaebel, W., & Wölwer, W. (1992). Facial expression and emotional face recognition in schizophrenia and depression. *European Archives of Psychiatry and Clinical Neuroscience, 242,* 46–52. http://dx.doi.org/10.1007/BF02190342

George, M. S., Nahas, Z., Molloy, M., Speer, A. M., Oliver, N. C., Li, X.-B., . . . Ballenger, J. C. (2000). A controlled trial of daily left prefrontal cortex TMS for treating depression. *Biological Psychiatry, 48,* 962–970. http://dx.doi.org/10.1016/S0006-3223(00)01048-9

Girard, J. M., & Cohn, J. F. (2015). Automated audiovisual depression analysis. *Current Opinion in Psychology, 4,* 75–79. http://dx.doi.org/10.1016/j.copsyc.2014.12.010

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., . . . Morency, L.-P. (2014). The Distress Analysis Interview Corpus of human and computer interviews. *Proceedings of Language Resources and Evaluation Conference*, 3123–3128.

Hamm, J., Kohler, C. G., Gur, R. C., & Verma, R. (2011). Automated Facial Action Coding System for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of Neuroscience Methods, 200,* 237–256. http://dx.doi.org/10.1016/j.jneumeth.2011.06.023

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *ArXiv:1512.03385 [Cs. CV].*

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9,* 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 17,* 299–310. http://dx.doi.org/10.1109/TKDE.2005.50

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16,* 606–613. http://dx.doi.org/10.1046/j.1525-1497.2001.016009606.x

Kuny, S., & Stassen, H. H. (1993). Speaking behavior and voice sound characteristics in depressive patients during recovery. *Journal of Psychiatric Research, 27,* 289–307. http://dx.doi.org/10.1016/0022-3956(93)90040-9

Mallett, S., Halligan, S., Thompson, M., Collins, G. S., & Altman, D. G. (2012). Interpreting diagnostic accuracy studies for patient care. *British Medical Journal (Clinical Research Ed.), 345,* e3999. http://dx.doi.org/10.1136/bmj.e3999

Marsland, S. (2011). *Machine learning: An algorithmic perspective*. New York, NY: Taylor & Francis. http://dx.doi.org/10.1201/9781420067194

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing, 3,* 5–17. http://dx.doi.org/10.1109/T-AFFC.2011.20

Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., & Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *Journal of Neurolinguistics, 20,* 50–64. http://dx.doi.org/10.1016/j.jneuroling.2006.04.001

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Petricoin, E. F., & Liotta, L. A. (2004). SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. *Current Opinion in Biotechnology, 15,* 24–30. http://dx.doi.org/10.1016/j.copbio.2004.01.005

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (pp. 1–8). Shanghai, China.

Salvatore, C., Cerasa, A., Castiglioni, I., Gallivanone, F., Augimeri, A., Lopez, M., . . . Quattrone, A. (2014). Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and progressive supranuclear palsy. *Journal of Neuroscience Methods, 222,* 230–237. http://dx.doi.org/10.1016/j.jneumeth.2013.11.016

Schuller, B., Steidl, S., & Batliner, A. (2009). The interspeech 2009 emotion challenge. *10th Annual Conference of the International Speech Communication Association* (pp. 312–315). Brighton, United Kingdom.

Substance Abuse and Mental Health Services Administration. (2017). Key substance use and mental health indicators in the United States: Results from the 2016 National Survey on Drug Use and Health (HHS Publication No. SMA 17–5044, NSDUH Series H-52). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., & Krajewski, J., Pantic, M. (2014). Avec 2014: 3rd dimensional affect and depression recognition challenge. *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* (pp. 3–10). Orlando, Florida.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., . . . Pantic, M. (2013). AVEC 2013: The continuous audio/visual emotion and depression recognition challenge. *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge* (pp. 3–10). Barcelona, Spain.

Wang, J., & Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *ArXiv:1712.04621 [Cs. CV]*.

Wang, P., Barrett, F., Martin, E., Milonova, M., Gur, R. E., Gur, R. C., . . . Verma, R. (2008). Automated video-based facial expression analysis of neuropsychiatric disorders. *Journal of Neuroscience Methods, 168,* 224–238. http://dx.doi.org/10.1016/j.jneumeth.2007.09.030

World Health Organization. (2017). Depression and other common mental disorders. (WHO reference number: WHO/MSD/MER/2017.2). Geneva, Switzerland: World Health Organization.

Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M. C., & Sahli, H. (2017). Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge-AVEC '17* (pp. 53–59). New York, NY: ACM Press.

Yang, Y., Fairbairn, C., & Cohn, J. F. (2013). Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing, 4,* 142–150. http://dx.doi.org/10.1109/T-AFFC.2012.38

Yao, Y., Rosasco, L., & Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation, 26,* 289–315. http://dx.doi.org/10.1007/s00365-006-0663-2

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *ArXiv:1506.06579 [Cs. CV]*.

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. *ArXiv:1702.04595 [Cs. CV]*.

Appendix

Details of the Human-Computer Interactive Evaluation

The first eight questions that participants responded to were related to their current mental well-being and were included for model training. These questions were as follows: (a) How have you been feeling lately? (b) Tell me how your sleep has been lately. (c) What else is going on today? (d) What are you looking forward to in the near future? (e) What's been frustrating you lately? (f) What do you think is causing your problems? (g) How would you describe the impact your life has on the world around you? (h) Whom do you wish you had a better relationship with, and what would make it better?

Participants also responded to five additional questions regarding past and current treatment history: (a) Are you currently treating depression with a mental health professional? (b) Are you currently treating anxiety with a mental health professional? (c) Has a mental health professional diagnosed you with depression in the past? (d) Has a mental health professional diagnosed you with anxiety in the past? and (e) Have you ever been treated for substance abuse or dependence?

Ideal recording conditions for participants were defined as such: (a) face within the dashed green lines (safe area), (b) no facial accessories (hats, glasses, etc.), (c) participant directly facing the camera, (d) participant's entire head in the frame, (e) face well lit from the front, and (f) minimized background distractions and noise. They were also asked to avoid backlighting and facing the camera at an angle.

Received October 1, 2018
Revision received March 8, 2019
Accepted March 17, 2019 ■