The Quest of Cost-effective Models Detecting Depression from Speech

Anonymous Author(s)

Affiliation
Address
email

Abstract

In this work, we explore the effectiveness of two different acoustic feature groups - conventional hand-curated and deep representation features, for predicting the severity of depression from speech. We measure the relevance of possible contributing factors to the models' performance, including gender of the individual, severity of the disorder, content and length of speech. Our findings suggest that models trained on conventional acoustic features perform equally well or better than the ones trained on deep representation features at significantly lower computational cost, irrespective of other factors, e.g. content and length of speech, gender of the speaker and severity of the disorder. This makes such models a better fit for deployment where availability of computational resources is restricted, such as real time depression monitoring applications in smart devices.

1 Introduction

Depression is a common psychological disorder affecting about 300 million people worldwide [Depression Fact Sheet, WHO]. Conventional depression diagnostic systems, such as clinical assessment or standard questionnaires, require significant amount of time and active participation of the depressed individuals. Studies reveal that depression is reflected in behavioral fluctuations of certain day-to-day activities and physical parameters [Wang et al., 2014], among which audiovisual is one of the most explored areas. In this work, we emphasize on audio modality for its manifold benefits. Audio based depression detection system offers better privacy for users of remote monitoring system. This kind of automated assessment takes only a few minutes of audio recording, therefore is less time-consuming, and would reduce burden on the individuals.

Existing best performing machine learning (ML) models that detect mental and cognitive diseases from audio data use either deep representation acoustic features, or a combination of conventional hand-crafted and deep features [Ray et al., 2019]. Although deep representation features offer a unified process of feature extraction, feature selection and model training, extracting and processing these features demands enormous computation resources including memory and processing time. This makes such models inconvenient for many real-world applications, where speed of data processing, model training and inference are of crucial importance [Yalamanchili et al., 2020]. Therefore, researchers and system designers need to make a choice of features when developing and deploying the model, considering both performance and cost. Some previous research compare the two approaches in the domain of cognitive disease detection [Balagopalan and Novikova, 2021] but to the best of our knowledge, no such research has been done so far in the domain of depression. To address this gap, in this work we have experimented with both hand-crafted conventional acoustic features and deep representation acoustic features. We address the following research questions: (1) Between conventional and deep representation acoustic features, which ones are more effective in determining depression severity in terms of accuracy and computational cost? (2) Does the machine...

In this work, we compare performance of the ML models trained on each type of features extracted from speech samples of a variety of content and length. Our key findings suggest that: (1) ML model trained on conventional acoustic feature set curated using expert domain knowledge demonstrates competitive performance as state-of-the-art models in predicting depression severity, irrespective of length and content of speech, and gender of the speaker, and (2) Usage of deep representation features resulted in marginal improvement of performance (0.0004%) consuming 1000 times more memory and 3000 times more computation time.

2 Materials and Methods

2.1 Dataset

We use DEPression and Anxiety Crowdsourced corpus (DEPAC) [Tasnim et al., 2022] in this experiment. The dataset consists of 2,674 audio samples collected from 571 English-speaking subjects, recording five self-administered speech tasks including two guided phonemic tasks and three free form speech tasks. The average length of the speech samples ranged from 5 to 50 seconds. In this dataset, the depression severity is measured in PHQ-8 scale (range: 0 to 24). The mean PHQ-8 score of DEPAC corpus (M) is 6.56 with standard deviation (SD) of 5.56. More details on DEPAC dataset is in [Tasnim et al., 2022].

2.2 Audio Quality Enhancement

To suppress possible background noise present in the samples and improve quality of the audio, we applied logmmse enhancement technique [Ephraim and Malah, 1985] on the audio samples. This method was found the best among existing audio enhancement algorithms [Hu and Loizou, 2006]. The enhancement step is found statistically significant \( p \leq 0.005 \) on 94% of the 220 conventional acoustic features in Wilcoxon signed-rank test with Bonferroni correction. Audio volume was normalized to -20 dBFS across all speech segments to control for variation caused by recording conditions such as microphone placement.

2.3 Acoustic Features

We extracted two sets of acoustic features, representing hand-crafted sets of conventional features and deep learning features:

**Conventional Acoustic Features:** This set included 220 acoustic features, extracted from each audio sample. The feature set includes spectral features e.g. intensity (auditory model based), MFCC 0-12, zero-Crossing Rate (ZCR), and voicing-related features, such as fundamental frequency \( (F_0) \), Harmonic-to-Noise Ratio (HNR), shimmer and jitter. Statistical functionals including minimum, maximum, average, and variance were computed on the low-level descriptors. Additionally, skewness and kurtosis were calculated on MFCCs, first and second order derivatives of MFCCs, and Zero Crossing Rate (ZCR) [Low et al., 2020].

**Deep Representation Features:** We used DeepSpectrum library [Amiriparian et al., 2017] to extract features from a pre-trained VGG-16 Convolutional Neural Network (CNN) [Simonyan and Zisserman, 2014]. The speech files are first transformed into mel-spectrogram images with 128 mel-frequency bands. Then, the spectral images are forwarded through the pre-trained networks. A 4,096-dimensional feature vector is then formed from the activations of the second fully connected layer in VGG-16. The features were extracted at a window width of 1s and a hop size of 300 ms from each audio sample.

2.4 Data Preprocessing

**Standardization:** The range of values of audio features tends to vary widely. To ensure even contribution of all features in the regression task, and to speed up gradient descent convergence of the deep neural network, once acoustic features were extracted from the audio samples, we standardized them using z-scores, i.e., subtracting the mean and dividing by standard deviation.
2.5 Model Training

We train Support Vector Machines (SVM), Random Forest (RF) and Feedforward Neural Network (FNN) as representative of linear and non-linear ML models. We train the models separately on conventional and deep learning acoustic features (Details in Appendix C).

We present evaluation metrics with 5-fold cross-validation (CV) for the models. The folds are subject-independent, i.e. samples from any subject does not appear in both training and test set. The folds preserve the same ratio of depression severity in each training and test partitions.

To understand the effect of speech content on ML models’ performance, we separated samples with each type of speech task and trained models on each type of them. We repeated the same process for conventional and VGG-16 features.

We ran our experiment on a MacBook Pro with Intel Core i7 CPU at clock speed of 2.67 GHz. The system availed 16 GB memory. The data preprocessing and model training was done in Python programming language.

<table>
<thead>
<tr>
<th>Gender</th>
<th>SVM RMSE</th>
<th>SVM MAE</th>
<th>RF RMSE</th>
<th>RF MAE</th>
<th>FNN RMSE</th>
<th>FNN MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv.</td>
<td>VGG-16</td>
<td>Conv.</td>
<td>VGG-16</td>
<td>Conv.</td>
<td>VGG-16</td>
<td>Conv.</td>
</tr>
<tr>
<td>Male</td>
<td>5.04</td>
<td>7.89</td>
<td>4.22</td>
<td>6.95</td>
<td>5.15</td>
<td>5.06</td>
</tr>
<tr>
<td>Female</td>
<td>5.64</td>
<td>7.11</td>
<td>4.33</td>
<td>6.23</td>
<td>5.47</td>
<td>5.51</td>
</tr>
<tr>
<td>Overall</td>
<td>5.38</td>
<td>7.48</td>
<td>4.28</td>
<td>6.56</td>
<td>5.32</td>
<td>5.31</td>
</tr>
</tbody>
</table>

3 Result and Discussion

3.1 Performance of models trained on conventional and deep representation features

We trained 3 different ML models on each type of acoustic feature, i.e. conventional and VGG-16. We report the RMSE and MAE error of each model trained separately on samples from male and female subjects, along with the overall performance on the entire dataset (Table 1).

SVM and FNN models performed better on conventional features than on VGG-16, while performance of RF is marginally better (0.0004%) on VGG-16 (Table 1). These findings are consistent with the previous works (Appendix D Table 1). In comparison to the state of the art acoustic models, our proposed RF models show competitive performance. The RF model trained on both types of features outperforms almost all the existing works reporting similar performance metrics on PHQ-8 scale. Only Ray et al. [2019] reported lower RMSE than us, using a combination of conventional and deep representation features, and formulating a multi-level LSTM architecture. Our proposed RF model trained on conventional features produces competitive performance to their proposed model, while substantially decreasing computational requirements.

The VGG-16 features collected in the same manner as described by Ray et al. [2019] from our audio corpus occupy 11.21 GB of memory, while our presented conventional feature file size is only 11 MB. Preprocessing and training models on conventional acoustic features took on average 3 minutes, while the procedure on VGG-16 features took at least 150 hours on the same computational environment (2.6 GHz 6 core Intel Core i7 processor, 16 GB memory). In short, our RF model using VGG-16 features offers 0.0004% improvement in performance than the same model using conventional features, using 1000 times more memory and 3000 times more processing time, implying similar or more computational resource is required for training complex models on multimodal features for marginal performance improvement. Therefore, the conventional features provided better opportunity to adjust model parameters for performance improvement.
In most cases, RMSE and MAE are lower for male subjects than female subjects. The reason behind this can be the lower severity of depression among male subjects than females in DEPAC dataset [Tasnim et al., 2022]. The skewness in the dataset causes bias in model prediction, as described by Larrazabal et al. [2020]. For real world applications, this issue needs to be taken care of by ensuring gender balance in training data.

Comparison of the performance of our best model with the state-of-the-art is in Table 3 (Appendix D). CPU time required to train each model is summarized in Table 4 (Appendix D).

We performed two-sample t-tests to identify if the performance deviations of the models are significant when trained on conventional acoustic features and VGG-16 features. There was a significant difference between absolute errors in predictions of SVM models trained on conventional acoustic features (M = 4.28, SD = 3.25) and VGG-16 features (M = 6.5, SD = 3.59); t(5332) = -24.20, p = 6.86e-123 < .05. The absolute errors in predictions of SVM with model is significantly higher when trained on VGG-16 features than when trained on conventional features. On the other hand, there was no significant difference between absolute errors in predictions of our best performing RF and FNN models trained on conventional acoustic features (RF : M = 4.34, SD = 3.09; FNN : M = 4.32, SD = 3.14) and VGG-16 features (RF : M = 4.31, SD = 3.11; FNN : M = 4.48, SD = 3.11).

### 3.2 Effect of speech task type, speech length and depression severity on ML model performance

We did not observe any significant deviation of model performance on the basis of speech task (See Table 5, Appendix D), therefore it is possible to recommend as a design choice any speech task of a similar length and content.

No significant correlation is found between model performance (absolute error of each prediction) and length of corresponding sample. The Concordance Correlation Coefficient (CCC) score for SVM, RF and FNN models are 0.000, 0.003 and -0.002 for conventional features and -0.004, -0.006 and -0.010 for VGG-16 features respectively. The near-zero CCC values indicate that in the case of our dataset, the speech length of samples does not influence the models’ performance (See visualization in Figure 1 and 2 Appendix D). Note that all speech samples in DEPAC are less than one minute.

The CCC score for SVM, RF and FNN models are 0.055, 0.385 and 0.435 for conventional features and -0.726, 0.445 and 0.394 for VGG-16 features respectively. The high positive CCC scores for most of the models imply that the samples with higher PHQ-8 scores contribute more to the overall prediction error of the models. This is caused by the imbalance in the number of samples with high and low PHQ-8 scores in DEPAC dataset [Tasnim et al., 2022]. The higher density of samples with subthreshold (≤ 5) PHQ-8 score bias the models to make predictions close to the mean PHQ-8 (6.56) of the dataset. This observation strengthens the necessity of balancing the samples in training models to be used in real world application.

Absolute errors for each sample plotted against ground truth PHQ-8 score for the models trained on conventional and VGG-16 features can be found in Figure 3 and 4 Appendix D.

### 4 Conclusion and Future Works

Speech has proven to be a reliable marker for depression assessment. But in order to deploy a machine learning model in a practical system, it is necessary to identify the most informative acoustic feature, along with an efficient and cost-effective process to train the model. In this paper, we study the performance of conventional acoustic feature-based and pre-trained deep representation based models on predicting depression severity from speech. We observe that the hand-curated feature based approach achieves better performance at a remarkably less computation time. Our experiments show that gender of the speaker and distribution of score affect the model performance, and should be taken care of while formulating balanced training data. However, content and length of speech do not show significant impact as long as the length of speech samples is reasonably short, less than one minute in our case. In our future work, we plan to explore generalizability of the findings across other datasets and disorders.
References


Appendix

A Related Works

Individuals suffering from psychological and neurological disorders like depression exhibit measurable fluctuation in vocal parameters (Darby et al. [1984] and Cummins et al. [2014]). Significant number of research have been conducted to relate these parameters with depression severity. DAIC-WoZ dataset [Gratch et al., 2014] is a widely used dataset in acoustic based depression severity prediction, consisting of structured interviews of participants conducted by a virtual agent. Two subsets of this dataset have been introduced as the challenge corpus of three Audio/Visual Emotion Challenges (AVEC) in 2016 [Valstar et al., 2016], 2017 [Ringeval et al., 2017] and 2019 [Ringeval et al., 2019], where participants proposed machine learning models to predict depression score on the PHQ-8 scale [Kroenke et al., 2001]. Handcrafted acoustic features have been exploited for this task for the last few decades, while deep representation of acoustic features have become popular in recent years. Further, we present a summary of existing works in this area and compare them based on the type of acoustic features.

A.1 Conventional Acoustic Features

Conventional acoustic features fall in temporal, spectral, energy and voicing related categories, from which researchers hand-pick the ones that are most suitable for predicting certain disorders, such as depression [Cummins et al., 2014]. Over the time, certain sets of these acoustic features, introduced in speech emotion and depression recognition challenges, have gained popularity, among which baseline feature sets of AVEC 2013 [Valstar et al., 2013] and AVEC 2016 [Valstar et al., 2016], INTERSPEECH ComParE [Schuller et al., 2013], extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [Eyben, 2015] are noteworthy. Development of feature extraction toolkits like openSMILE [Eyben et al., 2010], COVAREP [Degottex et al., 2014] has made it easier for researchers to extract these features for the purpose of speech analysis in different aspects.

A.2 Deep Representation Features

Deep representation of acoustic features are inspired by the deep learning paradigms common in image processing. Here, spectral images of speech instances are fed into pre-trained image recognition CNNs and a set of the resulting activations are extracted as feature vectors. In AVEC 2019 Depression Detection Sub-challenge (DSC), Deep representation features from four robust pre-trained CNNs using VGG-16 [Huang et al., 2017], AlexNet [Krizhevsky et al., 2012], DenseNet-121, and DenseNet-201 [Simonyan and Zisserman, 2014] were included as challenge baseline features. Participants chose between using one or more sets of deep representation features (Rodrigues Makiuchi et al. [2019], Yin et al. [2019]), and combining them with traditional features [Ray et al., 2019] and obtained competitive performances (Table 3). Deep representation provides the option to unite feature extraction, feature selection and model training into a single automated generalisable procedure, compromising the opportunity to incorporate expert domain knowledge, and necessitating considerably higher computational cost.

A.3 Depression Detection Models

AVEC 2016 challenge dataset was used in analysis presented by Williamson et al. [2016], Sun et al. [2017], Gong and Poellabauer [2017], Yang et al. [2017b], Samareh et al. [2018], Al Hanai et al. [2018], Haque et al. [2018], Zhao et al. [2020] and Muzammel et al. [2020]. Williamson et al. [2016] extracted formants, MFCCs, glottal features, loudness. In addition to COVAREP [Degottex et al., 2014] audio features, Sun et al. [2017] took text topics into account, while Gong and Poellabauer [2017] considered a more extended set of features of audio, video and text modalities. Samareh et al. [2018] added Delta and Delta-Delta coefficients, mean, median, standard deviation, peak-magnitude to RMS ratio to the set of challenge baseline audio features. Applying similar higher-order statistics on the baseline features, Al Hanai et al. [2018] constructed an extended feature set of 553 features, of which they identified 279 features with statistically significant univariate correlation. Haque et al. [2018] implemented multi-modal sentence-level embedding on log-Mel spectrogram and MFCC features. In their recent work, Yang et al. [2020] exploited a combination of eGeMAPS and INTERSPEECH features extracted from the longest ten segments of each audio
sample. They reshape the feature vector in an image-like 2D feature map in row-major order and 
adopted Deep Convolutional Generative Adversarial Net (DCGAN) for feature vector generation. 
Muzammel et al. [2020] trained three spectrogram-based Deep Neural Network architectures phoneme 
consonant and vowel units and their fusion. Their findings suggest that deep learned consonant-based 
acoustic characteristics lead to better recognition results than vowel-based ones, and the fusion of 
vowel and consonant speech characteristics outperforms the other models on the task. Zhao et al. 
[2020] described a transfer attention mechanisms from speech recognition to aid depression severity 
measurement. The transfer is applied in a two-level hierarchical network which reflecting the natural 
hierarchical structure of speech.

The AVEC 2016 challenge corpus included training, development and test partitions of audio samples. 
Using acoustic features exclusively, the lowest root-mean-square-error (RMSE) of 5.52 and 6.42 on 
the development and test set were reported in [Yang et al., 2017a] and [Syed et al., 2017] respectively. 
RMSE 4.99, 5.40, 5.66 and 6.42 were reported by Gong and Poellabauer[2017], [Yang et al., 2017a], 
Zhao et al. [2020] and Syed et al. [2017] respectively on the test set. The challenge baseline RMSE 
6.74 (mean absolute error (MAE) 5.36) and 7.78 (MAE 5.72) were set for the development and test 
set respectively [Ringeval et al., 2017].

Comparatively, fewer number of studies have been conducted on the recently published AVEC 2019 
DDS dataset; which is a super-set of the AVEC 2016 dataset. A wide range of audio features has been 
provided as the challenge baseline encompassing both handcrafted sets of conventional features and 
deep representation of acoustic features, including eGeMAPS, Mel Frequency Cepstral Coefficients 
(MFCCs), Bag of Audio Words (BoAW) and two sets of deep spectrum features by feeding spectral 
images of speech instances into pre-trained image recognition Convolutional Neural Networks (CNN) 
(VGG-16 [Simonyan and Zisserman 2014] and DenseNet-121 [Huang et al. 2017]) and extracting 
the resulting activations as feature vectors. Of these, deep spectrum features were used by Yin et al. 
[2019] and Rodrigues Makuchi et al. [2019], and MFCCs and eGeMAPS by Fan et al. [2019]. Ray 
et al. [2019] exploited all the baseline feature sets, while Zhang et al. [2020] extracted the AVEC 
2017 baseline feature set [Ringeval et al. 2017] using the COVAREP software toolbox [Degottex 
et al. 2014], in addition to eGeMAPS. Different configurations of long short-term memory (LSTM) 
networks were used in all of these works except [Zhang et al. 2020], who adopted random forest and 
logistic regression. Acoustic models proposed by [Ray et al. 2019] and [Zhang et al. 2020] achieved 
lowest RMSE of 5.11 and 6.78 on the development and test partitions, respectively.

B Dataset Description

The DEPAC corpus consists of 2,674 audio samples collected from 571 subjects located in Canada 
and the United States. 54.67% of the study subjects are female and 45.33% are male, aged between 
18 and 76 years, and they received 1 to 26 years of formal education. The data was collected via 
crowdsourcing and consists of a variety of self-administered speech tasks (Table 2). The participants 
completed these tasks using Amazon Mechanical Turk (mTurk). The speech tasks were curated to 
increase phonemic variety and were supported by literature on detecting mental disorders, such as 
Alzheimer’s Disease (AD) [Borkowski et al. 1967] and depression [Jiang et al. 2017], [Fossati et al. 
2003], [Trifu et al. 2017] from speech.

In this dataset, the depression severity is represented by Patient Health Questionnaire (PHQ-9) 
scores. It is a 3-point self-rated measure for depressive symptoms, including 9 questions. To ensure 
comparability of our results with works done on popular subsets of DAIC-WoZ corpus [Gratch et al. 
2014], i.e. AVEC 2017 [Ringeval et al. 2017] and AVEC 2019 [Ringeval et al. 2019], we used 
responses to 8 PHQ questions in our analysis and reported our results on PHQ-8 scores. The score 
ranges from 0 to 24 on PHQ-8 scale where a score higher than 5, 9 and 14 represent mild, moderate 
and severe level of depression respectively. The mean PHQ-8 score of DEPAC corpus (M) is 6.56 
with standard deviation (SD) of 5.56.

C Depression Models

Following Balagopalan and Novikova [2021] and Tasnim and Stroulia [2019], we train a combination 
of linear and non-linear ML models separately on conventional and deep learning acoustic features:

https://www.mturk.com
Table 2: Speech tasks in DEPAC corpus

<table>
<thead>
<tr>
<th>Speech Task</th>
<th>Description</th>
<th>Average Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phoneme Task</td>
<td>Record “aah” sound for as long as the participant could hold breath</td>
<td>5.79 sec</td>
</tr>
<tr>
<td>Phonemic Fluency</td>
<td>Pronounce as many unique words as possible starting with the letters “F”, “A” or “S”</td>
<td>22.13 sec</td>
</tr>
<tr>
<td>Picture Description</td>
<td>Describe a picture shown on the screen</td>
<td>46.60 sec</td>
</tr>
<tr>
<td>Semantic Fluency</td>
<td>Describe a positive experience they expected to have within five years in future</td>
<td>43.76 sec</td>
</tr>
<tr>
<td>Prompted Narrative</td>
<td>Tell a personal story, describing the day, a hobby, or a travel experience</td>
<td>45.34 sec</td>
</tr>
</tbody>
</table>

(a) SVM (CCC = 0.000)  
(b) RF (CCC = 0.003)   
(c) FNN (CCC = -0.002)

Figure 1: Correlation between speech length and prediction error of models trained on conventional acoustic features

- Support Vector Machines (SVM): Radial Basis Function (RBF) kernel SVM was trained. Values of hyperparameters ‘C’ and ‘gamma’ were tuned by 5-fold grid-search cross validation (cv).
- Random Forest (RF): Scikit Learn implementation of random forest regressor was used. Number of estimator trees and maximum depth were tuned through grid-search cv.
- Feedforward Neural Network (FNN): The FNN model consists of 4 hidden layers, with 500 hidden units on the first layer, 250 in the second and 125 in the rest of the hidden layers. 30% dropout on the output of each of the hidden layers of the FNN. We use Adam optimizer in all FNN models with the learning rate of 0.001. Each of the FNN models is trained for 150 epochs.

The discussion presented by Balagopalan and Novikova [2021] suggest that for small audio corpus like the ADReSSo challenge dataset (237 samples) [Luz et al., 2021], either leave-one-subject-out cross validation or k-fold cross validation can be applied. However, the dataset used in this work is considerably larger than the ADReSSo challenge dataset. Considering the size of the dataset and corresponding computational complexity, we decided to report evaluation metrics with 5-fold cross-validation (CV) for the models. We create 5 subject-independent folds, train the model using 4 of them, and use the rest for testing. We repeat the process for all 5 folds and report evaluation metrics averaging across predictions on all the folds. These folds preserve the same ratio of depression severity in each training and test partitions.

**D Model Performance**
Table 3: Comparison of performance of SOTA ML models trained on different combinations of features. Bold denotes regression error of our proposed model.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Study</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>Formants, MFCCs, glottal features, loudness, AVEC 2017 dataset</td>
<td>6.38</td>
<td>5.32</td>
</tr>
<tr>
<td></td>
<td>[Williamson et al., 2013]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>COVAREP feature set, AVEC 2017 dataset [Syed et al., 2017]</td>
<td>6.34</td>
<td>5.30</td>
</tr>
<tr>
<td></td>
<td>COVAREP features and functional, AVEC 2016 dataset [Al Hanai et al., 2018]</td>
<td>6.50</td>
<td>5.13</td>
</tr>
<tr>
<td></td>
<td>MFCC, AVEC 2016 dataset [Haque et al., 2018]</td>
<td>5.78</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>MFCC and eGeMAPS features, AVEC 2019 dataset [Fan et al., 2019]</td>
<td>6.20</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>eGeMAPS, INTERSPEECH features, AVEC 2016 dataset [Yang et al., 2020]</td>
<td>5.52</td>
<td>4.63</td>
</tr>
<tr>
<td></td>
<td>eGeMAPS and COVAREP features, AVEC 2019 dataset [Zhang et al., 2020]</td>
<td>6.78</td>
<td>5.77</td>
</tr>
<tr>
<td>Deep representation</td>
<td>VGG-16 features, AVEC 2019 dataset [Rodrigues Makiuchi et al., 2019]</td>
<td>5.70</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Mel-spectra, AVEC 2017 dataset [Zhao et al., 2020]</td>
<td>5.66</td>
<td>4.28</td>
</tr>
<tr>
<td>Conventional deep</td>
<td>MFCC, BoAW, eGeMAPS and VGG-16 features, AVEC 2019 dataset [Ray et al., 2019]</td>
<td>5.11</td>
<td>-</td>
</tr>
<tr>
<td>combined</td>
<td>Conventional</td>
<td>MFCCs, HNR, jitter, shimmer, ZCR features, DEPAC dataset [Tasnim et al., 2022]</td>
<td>5.31</td>
</tr>
</tbody>
</table>

Table 4: Time elapsed in different stages of model training

<table>
<thead>
<tr>
<th>Processing step</th>
<th>Algorithm</th>
<th>Conventional</th>
<th>VGG-16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data loading</td>
<td>-</td>
<td>1.132</td>
<td>244450</td>
</tr>
<tr>
<td>Preprocessing</td>
<td>-</td>
<td>96.834</td>
<td>545483</td>
</tr>
<tr>
<td>Model training</td>
<td>SVM</td>
<td>1.600</td>
<td>5593</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.715</td>
<td>981</td>
</tr>
<tr>
<td></td>
<td>FNN</td>
<td>220.853</td>
<td>10270</td>
</tr>
<tr>
<td>Prediction</td>
<td>SVM</td>
<td>0.412</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.040</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>FNN</td>
<td>1.102</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>SVM</td>
<td>99.978</td>
<td>795579</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>98.715</td>
<td>790923</td>
</tr>
<tr>
<td></td>
<td>FNN</td>
<td>267.227</td>
<td>800210</td>
</tr>
</tbody>
</table>

(a) SVM (CCC = -0.004)  (b) RF (CCC = -0.006)  (c) FNN (CCC = -0.010)

Figure 2: Correlation between speech length and prediction error of models trained on VGG-16 features
Table 5: Regression error of models trained on speech samples of different tasks

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Speech task</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Conventional</td>
<td>VGG-16</td>
</tr>
<tr>
<td>SVM</td>
<td>Semantic fluency</td>
<td>5.30</td>
<td>6.62</td>
</tr>
<tr>
<td></td>
<td>Prompted narrative</td>
<td>5.29</td>
<td>6.63</td>
</tr>
<tr>
<td></td>
<td>Phoneme task</td>
<td>5.49</td>
<td>6.40</td>
</tr>
<tr>
<td></td>
<td>Phonemic fluency</td>
<td>5.45</td>
<td>6.49</td>
</tr>
<tr>
<td></td>
<td>Picture description</td>
<td>5.43</td>
<td>6.54</td>
</tr>
<tr>
<td>RF</td>
<td>Semantic fluency</td>
<td>5.24</td>
<td>5.24</td>
</tr>
<tr>
<td></td>
<td>Prompted narrative</td>
<td>5.31</td>
<td>5.25</td>
</tr>
<tr>
<td></td>
<td>Phoneme task</td>
<td>5.39</td>
<td>5.29</td>
</tr>
<tr>
<td></td>
<td>Phonemic fluency</td>
<td>5.38</td>
<td>5.29</td>
</tr>
<tr>
<td></td>
<td>Picture description</td>
<td>5.42</td>
<td>5.31</td>
</tr>
<tr>
<td>FNN</td>
<td>Semantic fluency</td>
<td>5.34</td>
<td>7.13</td>
</tr>
<tr>
<td></td>
<td>Prompted narrative</td>
<td>5.30</td>
<td>7.13</td>
</tr>
<tr>
<td></td>
<td>Phoneme task</td>
<td>5.49</td>
<td>7.33</td>
</tr>
<tr>
<td></td>
<td>Phonemic fluency</td>
<td>5.39</td>
<td>7.30</td>
</tr>
<tr>
<td></td>
<td>Picture description</td>
<td>5.45</td>
<td>7.29</td>
</tr>
</tbody>
</table>

Figure 3: Correlation between depression severity and prediction error of models trained on conventional acoustic features

(a) SVM (CCC = 0.055)  
(b) RF (CCC = 0.385)  
(c) FNN (CCC = 0.435)

Figure 4: Correlation between depression severity and prediction error of models trained on VGG-16 features

(a) SVM (CCC = -0.726)  
(b) RF (CCC = 0.445)  
(c) FNN (CCC = 0.394)