EmpathNet: Exploring Depression Detection from Semi-Clinical Interviews using Deep Learning

Pranav Vijay Chakilam
pchakila@iu.edu

Dharani Doppalapudi
ddoppala@iu.edu

Harsha Renkila
hrenkila@iu.edu

Abstract

Research in non-intrusive and convenient monitoring of depression levels, especially using deep learning techniques, has gained a lot of steam in recent years. In this paper, we aim to explore deep learning methods to predict depression based on video and audio data of individuals. A comprehensive literature survey is performed, mainly focusing on the various video and audio based approaches to solve this problem. The Extended Distress Analysis Interview Corpus database is used for this purpose. The Multiscale Spatiotemporal Network proposed in De Melo et al. [24] is selected as an effective approach for only video-based depression detection and its architecture is reimplemented in PyTorch. We employ and observe the efficacy of data augmentation techniques based on Gaussian Noise Injection and Google Brain's SpecAugment strategy. We experiment with Siamese networks for depression classification for the first time in the domain, to the best of our knowledge. Thirty-four models from the Keras Applications library are experimented with along with a simple 6-layer Custom CNN architecture, investigating the performance of simple and complex models for this problem. Finally, we investigate the effect of using text along with audio for the purpose of depression score prediction (regression). The maximum depression classification recall and accuracy achieved are 84% (by the Custom CNN Model) and 81% (by the Xception Model) respectively. A VGG16 model that uses both audio, and text encoded using Google's BERT, achieves the best performance while maintaining the ability to generalize over all PHQ-8 score ranges, resulting in a mean absolute error of 6.54.

2. Introduction

More than 300 million people of all age groups worldwide are estimated to suffer from depression [1]. In today's modern, hectic and stressful world, people are becoming increasingly depressed, miserable and unhappy even though our standards of living are improving. If not paid enough attention, our stressful lifestyles can lead to increasing anxiety which in turn results in depression which again increases our stress levels, severely deteriorating our mental health [2]. A simple, honest and effective solution to monitor, treat and prevent debilitating mental illnesses must be created to stop this silent pandemic. Today one of the most common ways to diagnose and monitor depression levels is to take surveys (for example - PHQ-8 survey) which can often get tedious for users. Various studies have shown that there exists a correlation between the mental health condition of an individual and his facial expressions, suggesting that depressed individuals are less likely to smile and more likely to exhibit gray expressions [3]. Apart from facial expressions the depression can be identified by the tone(speech) of the individuals and also from the meaning of the conversation(Text features). A simple computer vision powered solution combined with the speech features that can model these complex facial dynamic representations and dynamic audio representations can therefore aid in non-intrusively determining the depression score of an individual based on his/her facial videos, enabling more frequent and unobtrusive monitoring of their mental state. This can in turn help individuals understand their situation better and take proactive steps to prevent further worsening of their mental health.

3. Background and related work

In the past years, there has been an improved interest in unobtrusively detecting depression levels using Artificial Intelligence (AI), driving substantial academic and commercial research. Owing to the known efficacy of deep learning models to learn complex representations from facial and audio data, a significant amount of this research has focused on utilizing deep Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to predict depression levels from audiovisual cues. The Audio Visual Emotion Challenge (AVEC) 2013 and 2014 has contributed to this domain by serving as benchmark datasets by making raw visual and audio data available for research purposes. The datasets provided in the AVEC 2013 and AVEC 2014 challenges consisted of 150 and 300 videos respectively, where subjects performed tasks, such as reading aloud a passage or responding to questions such as "What is your sad childhood memory?". The Beck Depression Inventory-II (BDI-II) scores for the participating subjects are provided to serve as target labels for the corresponding videos.

One of the approaches explored in a vast amount of literature involves employing traditional schemes such as Motion History Histograms, Space-Time Interest Points, Pyramid of Histogram Gradients, Local Phase Quantization and Local Gabor Binary Patterns to extract features representations, while many methods utilize Machine Learning (ML) techniques like Support Vector Regression and Linear Regression Techniques to ascertain continuous depression levels [3, 4, 5, 6, 7, 8, 9, 10]. For example in Local Phase Quantization, it represents the multiresolution gray-scale and rotation invariant texture features of images with local binary patterns by creating the rotational invariant representation using the circular local neighborhood in binary pattern[6]. In [11], Valstar uses the RECOLA dataset and the FACET software in order to explain the statistics and outcomes for the AVEC 2016 dataset. However, these techniques do not take advantage of end-to-end deep learning and rely on engineered features that might often not capture the sophisticated facial dynamics involved in vision faced affective computing.

Subsequently, many single and multi-model deep learning-based approaches were proposed where CNNs and RNNs were taken advantage of to solve this problem. *Ranjan* [12], presented an overview of current advancements in creating an automated face recognition system and addressed several multitask CNN-based techniques for face analytics. *Ashraf* in [13], provided several insights on machine learning approaches and tools for creating detection models. In [10], *de Melo* models spatiotemporal information with the help of Convolutional 3D (C3D) networks. And in [14], *Al Jazaery* used an RNN-C3D formulation to learn the depression levels using the spatial and sequential temporal information in the video data. In Meng et al [15], CNN models are pre-trained by VGG- with a new Expectation loss in *Melo et al.* [19] for automated Face. A novel method called Feature Dynamic History Histogram (FDHH) is developed to capture temporal movement and the final mapping to depression scales is performed using Partial Least Squares, and Linear Regression techniques. Also, *Zhou* offered an effective strategy where a Depression Activation Map (DAM) is proposed and multi-region DepressNet, where multiple models are used for different regions of the face and their outputs are combined to improve the overall performance [16]. In [17], *Gavrilescu* implemented a model using CNNs to analyze facial expressions using the Facial Action Coding System. A ResNet model is trained along with depression detection. *Prabhudesai,* in [20], compared this ResNet + Expectation loss and few other models and analyzed the different frameworks (such as CNNs, RNNs, and others) and algorithms. *He,* in [21], compared *Zhou's* model on MR-DepressNet, *Melo's* RESNET + Expectation loss model with a Deep Local-Global Attention Convolutional Neural Network (DLGA-CNN), which uses DCNN with attention mechanism, and weighted spatial pyramid pooling that merged 2D-CNN networks with attention mechanism for identifying depression. In [22], *victor* created a system to predict if a person is depressed or not, which is AiME(Artificial Intelligence Mental Evaluation) which can operate with minimum human interference. Also, in [23], *Melo* employed a 3D CNNS model consisting of a parallel two-stream architecture that merges spatial and temporal information. In one of the most accurate works of depression assessment with facial dynamics [24], *Melo,* introduced a novel 3D CNN architecture and the Multiscale Spatiotemporal Network (MSN) to enable simultaneous analysis of the input videos using multiple parallel streams at different temporal ranges and receptive field sizes. To investigate the efficacy of the Knowledge Transfer (KT) technique for enabling deep learning on edge devices with low computational resources, *Sharma* in [25], a comprehensive study of four different KT techniques and three different model architectures is presented, evaluating their performance in terms of both accuracy and convergence time.

In contrast to traditional shallow features, *Xingchen Ma* in [28], proposed a deep learning model to capture depression characteristics in the voice channel, merging CNN and Long Short-Term Memory to deliver a fuller audio representation. He focused on sample imbalance and data representation. To balance the positive and negative samples, he employed a random sampling strategy in the model training phase. His averaging baseline model got 0.41 and the deep learning model got a better score than baseline model, of 0.52 F1 score. CNN and multipart interactive training were used by *Karol Chlasta*, in [29], to identify sadness in speech. Spectrograms are created automatically from audio samples that were used to apply data to residual CNNs. They achieved an accuracy of ResNet designs of about 77 percent. In [30], *Rodrigues Makiuchi* employed a Gated Convolutional Neural Network followed by an LSTM layer, as well as deep spectrum data collected from a trained VGG-16 network. For the textual embeddings, they employed an LSTM layer to extract BERT textual features. On the development set they received a CCC score of 0.696 and on the testing set of 0.403. End-to-end convolutional neural network models and spectrogram-based CNNs are used by *Srimadhur* in [31], for speech-based depression detection. Here, the end-to-end model outperformed the spectrogram-based model and baseline models by a factor of 13 percent. For landmark detection *Wu* [32], examined holistic approaches, Constrained Local Model methods, and regression-based methods. Holistic techniques include the Active Appearance Model and fitting algorithm. Face Shape Model and Local Appearance Model are examples of constrained local model approaches. A few regression-based approaches include direct regression, cascaded regression, and deep learning. According to the author, regression-based approaches surpass the other two for effective facial detection. For Databases, the author describes the "in-the-wild" database, which contains face data from various positions, illuminations, and occlusions. A recent facial detection technique termed "in the wild" facial landmark detection is also mentioned in this study. For landmark identification, this considers elements including head postures, occlusion, and face expressions.

Transcripts from video/audio, in addition to face video footage and audio samples, would aid in detecting the individual's mental state (sad, joyful, depressed, and so on) and help in enhancing depression detection. To identify the meaning of the text content we have a wide variety of techniques. BOW (Bag of words), TF-IDF (Term frequency and inverse document Frequency) are few naive methods that are used to extract the meaning of text content. But these are not so effective in extracting the meaning of text. There are few deep learning methods that convert words to vectors like Glove[36], Word2Vec, Fast text[37] and so on with RNN/LSTM context of the entire text. These deep learning methods provide an effective extraction of the meaning of text. A recent Language modeling technique by Google named BERT (Bidirectional Encoder Representation from Transformers) [33], is a state of art word embedding architecture. BERT is a self-attention-based transformer network that can encode the context of data in a meaningful fashion after being trained on a large corpus

4. Dataset and challenges

We found several publicly available datasets to detect depression, stress and anxiety from the face after a preliminary dataset. We aimed to concentrate on the AVEC dataset since it is the most well-known and widely used dataset with vast numbers of videos. We were first drawn to the Detecting Depression with AI Sub-challenge (DDS) of the AVEC 2019 dataset because it includes video footage from individuals as well as their PHQ-8 scores. However, after several trails and multiple efforts, we were eventually able to obtain access to the AVEC dataset of 2016.

We utilized the Extended Distress Analysis Interview Corpus (DAIC) dataset. This contains interviews with the goal of identifying a person's mental health by assessing their levels of depression, stress, and anxiety. These interviews are conducted by a virtual assistant named Ellie, who is controlled by a person in a separate room. This dataset contains 219 conversations with durations ranging from 7 to 33 minutes. The records of the interviews are kept in the form of audio, video, and long textual replies to the questions that were answered. The data include a variety of spoken and nonverbal features that have been transcribed and annotated. Within the dataset, the video material is divided into frames of images.

The most challenging hurdle we had along this path was acquiring access to the dataset so late and the dataset contains huge amounts of data. These impacted our timeline and we anticipated videos from the AVEC 2016 dataset, but instead we received image frames and audio samples. So, to quench our curiosity, we decided to try to apply computer vision on audio samples and see what the outcomes would be, to see whether applying Computer Vision to audio would have an influence and enhance accuracy. As a result, we had to adjust everything from detecting depression in facial footage to converting audio into spectrogram visuals. Few other challenges include the commonality of symptoms like similar pitches for depression, stress, anxiety and a variety of other mental illnesses. Therefore, it was difficult to differentiate and determine the depression levels. And performing data augmentation was tricky. We had to be extremely cautious while working with it as we weren’t able to use the usual color and position based data augmentation techniques such as flipping, rotating or adjusting the brightness for the spectrogram images like we do for other datasets because even a minor error may result in the loss of valuable underlying information.

5. Proposed Methods

- **Converting Audio Into Images:**

As stated in the previous section, we used the audio and text characteristics from the Extended Distress Analysis Interview Corpus (EDAIC) [43] to try to detect depression. Firstly, we transformed the audio samples to spectrogram images.

Example of a Mel-spectrograms is given below:

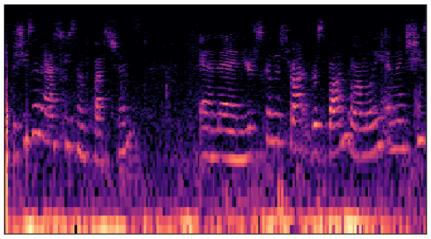


Fig 1: Spectrogram Images of Audio using mel frequency

The steps for conversion are as follows:

- ➢ There are different ways to extract the audio information but, we have picked Mel frequency as it best suits for human voice representation.
- ➢ For each individual interview the timeframe fluctuates, as the EDAIC dataset doesn't contain constant time audio interviews.
- ➢ Mel frequency spectrogram works for a variety of timeframes as it produces a consistent (288, 432, 3) picture regardless of audio duration.

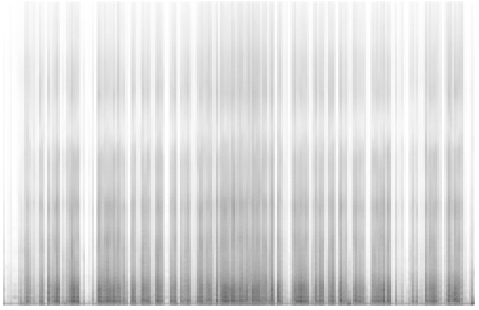


Fig 2: Spectrogram Images of Audio using MFCC

- ➢ The MFCC feature extraction[40] approach includes windowing the signal, using the DFT, obtaining the log of the magnitude, and then warping the frequencies on a Mel scale, followed by utilizing the inverse DCT. Mel spectrogram, on the other hand, is a transformation that shows the frequency composition of a signal over time[39].

Text characteristics encode utilizing the BERT:

- ➢ *BERT* model is a multi-head self-attention based transformer that can hold up to 512 tokens.
- ➢ The Bert tokenizer(approx.~= 33,000 inbuilt tokens) divides the words of a given sentence into tokens.

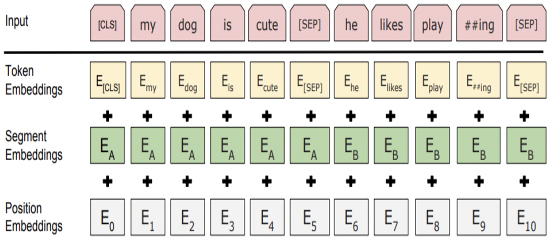

Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

Fig 3: The Bert Layer of Tokenisation[33]

- ➢ Encoded the representation of the given text of size 768 dimension vector using BERT.

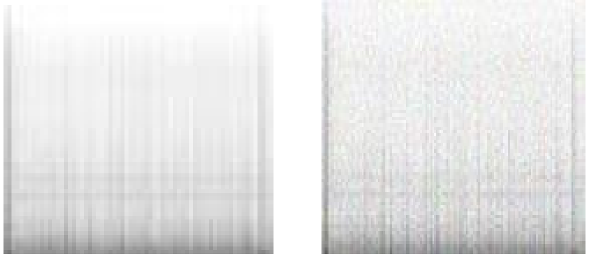
We explored the following strategies for detecting depression using the audio and text extraction methods described above.

- ➢ The Bert-large-Uncased model from the Sentence Transformers was utilized.
- ➢ They are mostly comparable variations of Bert, such as AlBert, DistillBert, and others, that have fewer parameters and behave similarly to Bert.

- **Data Augmentation:**

The dataset we had was substantially imbalanced with a majority (~77%) of the samples in the training data belonging to participants who were not depressed. Due to this, our initial Xception model learnt to predict ‘0’ i.e. not depressed or ‘1’ depressed for all the inputs it received. Therefore in order to solve this class imbalance problem, we considered 2 approaches. For the Xception model and the Custom CNN architecture used for depression classification, the number of depressed samples in the training were increased from around 33% to around 47% by performing data augmentation using 2 strategies - Gaussian Noise Injection, and SpecAugment strategy.

- ➢ **Gaussian Noise Injection**:

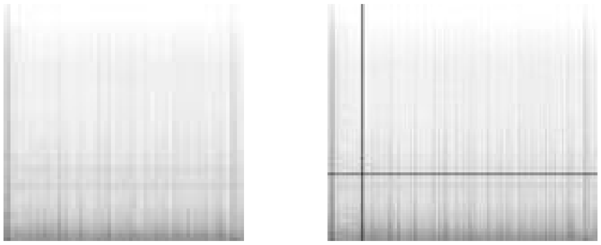


Original Image After Gaussian Noise Injection

Fig 4: Gaussian Noise Injection (Before and After)

Gaussian Noise injection adds white noise to the image, creating the pixelated effect on the left, we perform this using the wand python library.

- **SpecAugment Strategy:**



Original Image

After SpecAugment

Fig 5: SpecAugment Strategy (Before and After)

Another method we use for data augmentation is inspired from Google Brain's SpecAugment strategy. Here, we place random vertical and horizontal patches in the image masking segments in the frequency and time domains as shown above.

- **Approaches taken:**

We split our problem into 2 parts. Predicting whether a person has depression or not (classification) and predicting a person's PHQ-8 depression rating score (regression). The motivation behind this is the fact that regression, where our model has to predict a specific score, would obviously have a higher probability of being a more complex problem than just classifying individuals as depressed and not-depressed, given the small size of our dataset. Therefore, we take 2 approaches and explore separate techniques for both the problems as summarized in the diagram below (Fig 5). For classification, we experiment with a Siamese network, hoping to take advantage of their efficacy in situations where there is less data. We also consider 3 other CNNs that include the popular VGG16 and the Xception model along with our own Custom CNN architecture. For regression, we experiment with a VGG16 model that takes only audio as the input and another model that takes audio as well as text as the input.

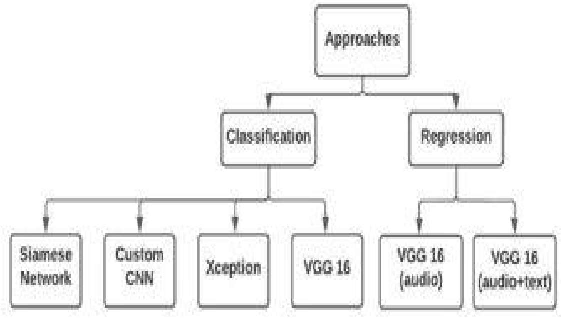


Fig 6: Approaches Taken.

- **Depression detection using only Audio:**

  - **Xception Model:**

In order to select a model for performing image classification in order to detect depression, we train 34 models present in the Keras Applications library on our dataset for 10 epochs. This is done to understand the potential of the models on the dataset. Then we sort the models based on their train and validation metrics (accuracy and loss). The table below shows the performances of the model sorted in ascending order based on their train and test accuracies. As we can see, the Xception model gives the best train and validation accuracy but does not have the maximum number of parameters.

Index	model name	no. of model params	validation accuracy	train accuracy
0	Xception	20861480	0.696428597	0.767295599
1	ResNet101V2	42626560	0.696428597	0.7547169924
2	MobileNet	3228864	0.696428597	0.7358490825
3	InceptionResNetV2	54336736	0.696428597	0.7232704163
4	ResNet152V2	58331648	0.696428597	0.716981113
5	VGG19	20024384	0.696428597	0.716981113
6	ResNet50	23564800	0.696428	0.710691

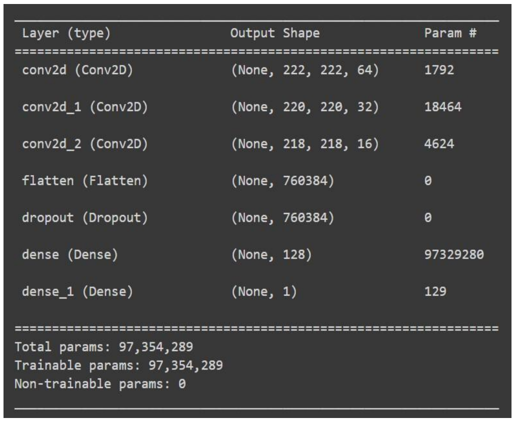
	V2		597	8097
7	EfficientNetB7	64097687	0.696428597	0.7044025064
8	InceptionV3	21802784	0.696428597	0.6918238997
9	DenseNet169	12642880	0.696428597	0.6855345964
10	EfficientNetV2L	117746848	0.696428597	0.6792452931
11	EfficientNetV2M	53150388	0.3035714328	0.6792452931
12	EfficientNetV2B1	6931124	0.696428597	0.6792452931
13	DenseNet121	7037504	0.6428571343	0.6792452931
14	MobileNetV2	2257984	0.696428597	0.6666666865
15	DenseNet201	18321984	0.4464285672	0.6666666865
16	NASNetMobile	4269716	0.696428597	0.6603773832
17	VGG16	14714688	0.696428597	0.6603773832
18	EfficientNetB0	4049571	0.696428597	0.6540880799
19	EfficientNetB3	10783535	0.696428597	0.647798717
20	MobileNetV3Large	2996352	0.696428597	0.6415094137
21	EfficientNetB2	7768569	0.696428597	0.6415094137
22	ResNet50	23587712	0.696428597	0.6415094137
23	EfficientNetV2B2	8769374	0.696428597	0.6352201104
24	EfficientNetB6	40960143	0.696428597	0.6289308071
25	EfficientNetB1	6575239	0.696428597	0.6289308071
26	EfficientNetV2S	20331360	0.696428597	0.6226415038
27	EfficientNetV2B0	5919312	0.696428597	0.6163522005
28	ResNet152	58370944	0.3035714328	0.6100628972
29	EfficientNetB4	17673823	0.696428597	0.6100628972
30	EfficientNetB5	28513527	0.696428597	0.6100628972
31	MobileNetV3Small	939120	0.696428597	0.6100628972
32	EfficientNetV2B3	12930622	0.696428597	0.6037735939
33	ResNet101	42658176	0.696428597	0.5597484112

Table 1: Model Selection Experiment Results on Keras Applications' models.

Therefore, we select the Xception model and fine tune it further, by training it on the augmented dataset, using the Adam Optimizer with a learning rate of 0.0001 and then further fine tuning it at an even smaller learning rate of 0.00001 and achieve respectable results shown in the results section. We hypothesize that the Inception philosophy employed in the Xception model allows it to analyze the image in various scales resulting in the impressive performance.

➢ Custom CNN Architecture:

In our experiments we noticed that although the model was achieving a good train and test performance the test set results especially the recall was relatively low for the Xception model. Therefore, we decided to experiment with a simple custom CNN architecture consisting of 3 2d convolutional layers followed by 2 fully connected (fc) layers, with the first layer using dropout. The architecture details are listed in Table 2 below. The results of the experiments performed with the model are listed in the results section.



```
Layer (type)                Output Shape              Param #
=================================================================
conv2d (Conv2D)             (None, 222, 222, 64)      1792

conv2d_1 (Conv2D)           (None, 220, 220, 32)      18464

conv2d_2 (Conv2D)           (None, 218, 218, 16)      4624

flatten (Flatten)           (None, 760384)            0

dropout (Dropout)           (None, 760384)            0

dense (Dense)               (None, 128)               97329280

dense_1 (Dense)             (None, 1)                 129

=================================================================
Total params: 97,354,289
Trainable params: 97,354,289
Non-trainable params: 0
```

Table 2: Custom CNN Architecture Outline

- **VGG16 Network[38]:**

The VGG16 network is replicated for our experiments having 5 VGG blocks with the 64,128,256,512,512 no of filters of size 3x3 for the layers respectively, followed by 4 fully connected layers of size 4096 → dropout → 1024 → dropout → 512 → 128 → 16 →1. This network is used for both regression and classification using pretrained weights and custom weights.

- **Depression detection using Audio and text:**

- As previously described, we wanted to combine audio and text features of an individual to assert the depression.
- We utilized VGG16 to encode the audio, then a pre-trained BERT network followed by dense layers to encode the text.
- While predicting the output (whether a person is depressed or not), the dense layers from both the pre-trained networks are concatenated.

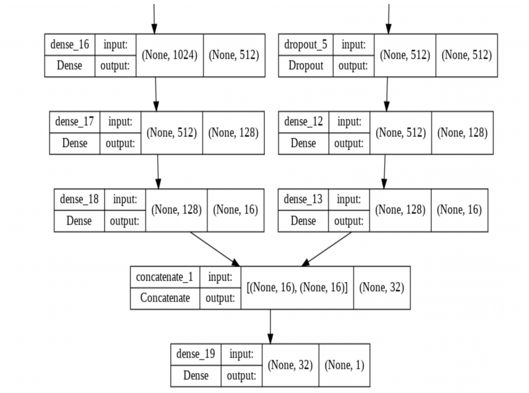

dense_16 | Dense | input: | output: | (None, 1024) | (None, 512)
dropout_5 | Dropout | input: | output: | (None, 512) | (None, 512)
dense_17 | Dense | input: | output: | (None, 512) | (None, 128)
dense_12 | Dense | input: | output: | (None, 512) | (None, 128)
dense_18 | Dense | input: | output: | (None, 128) | (None, 16)
dense_13 | Dense | input: | output: | (None, 128) | (None, 16)
concatenate_1 | Concatenate | input: | output: | [(None, 16), (None, 16)] | (None, 32)
dense_19 | Dense | input: | output: | (None, 32) | (None, 1)

Fig 7: Left - Dense layer followed by VGG16
Right - Dense layer from Bert network

- **Depression detection using Siamese Network:**

- Even for the less complex networks, deep neural networks require a vast quantity of data to produce the best outcomes.
- Because our dataset is so limited, we are utilizing spectrogram images instead of videos to train our deep neural networks.
- Siamese networks are widely renowned for their ability to handle limited data and class imbalances while maintaining high accuracy. Consider that there are 100 class A and 40 class B images. The new dataset will have 4000 image pairs. The label is 1 if both pairings are from the same class, otherwise it is 0.

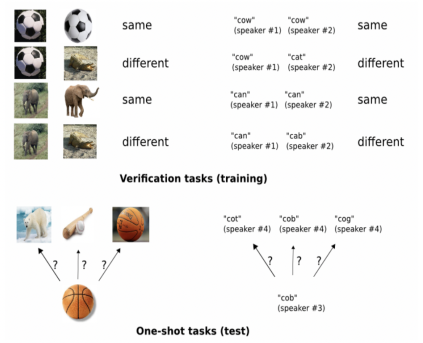

same
different
same
different
"cow" (speaker #1) "cow" (speaker #2) same
"cow" (speaker #1) "cat" (speaker #2) different
"can" (speaker #1) "can" (speaker #2) same
"can" (speaker #1) "cab" (speaker #2) different
Verification tasks (training)
"cot" (speaker #4) "cob" (speaker #4) "cog" (speaker #4)
? ? ?
? ? ?
"cob" (speaker #3)
One-shot tasks (test)

Fig 8: Siamese network image pair labeling[34]

- The network will take a pair of images as the input. so, the classification problem is modified into the similarity problem.
- The model has two networks, both of which are identical. Network will extract the features for both of them and calculate vector distances, and provide the similarity of both inputs as a training network.
- In these we have used contrastive Loss which calculates the similarity between two output vectors.

$$loss(d,Y)=\frac{1}{2}*Y*d^2+(1-Y)*\frac{1}{2}*\max(0,m-d)^2$$

- Where d is the distance between output of two networks, Y is the label, and m is the marginal parameter.

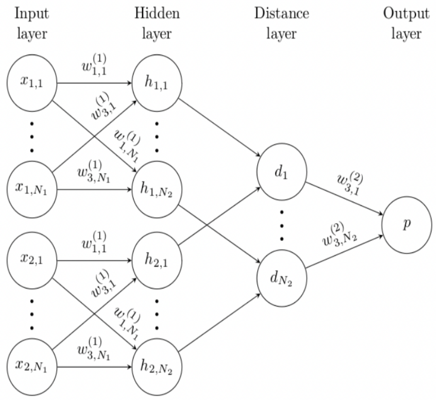


Fig 9: Siamese Network Architecture[34]

- ➢ One input is fixed during inference, and the query image is provided to identify the similarity of the query image to the fixed input and create the classification label.
- ➢ We used the non-pre-trained VGG16[38] as the basic model for our studies and attempted to detect depression using spectrogram images.

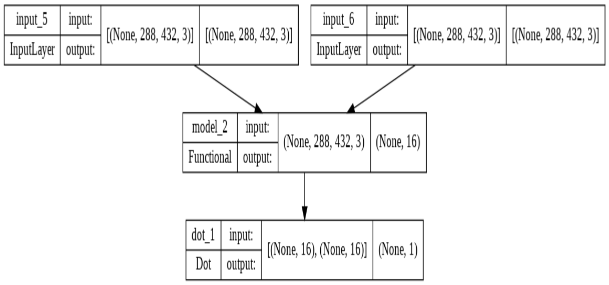

input_5 | input: | [(None, 288, 432, 3)] | [(None, 288, 432, 3)]
InputLayer | output:
input_6 | input: | [(None, 288, 432, 3)] | [(None, 288, 432, 3)]
InputLayer | output:
model_2 | input: | (None, 288, 432, 3) | (None, 16)
Functional | output:
dot_1 | input: | [(None, 16), (None, 16)] | (None, 1)
Dot | output:

Fig 10: Our Siamese network model as base model VGG16

A Literature Survey of 35 papers was conducted before selecting the most relevant ones to summarize in the related work section. The model proposed in [24], “A Deep Multiscale Spatiotemporal Network for Assessing Depression from Facial Dynamics” was identified as one of the best ones so far, to detect depression using facial videos.

As mentioned in the related work, this model uses an inception-style approach utilizing multiple parallel streams to analyze the input at different temporal ranges and spatial sizes, letting it capture complex dynamic representations (Fig. 7).

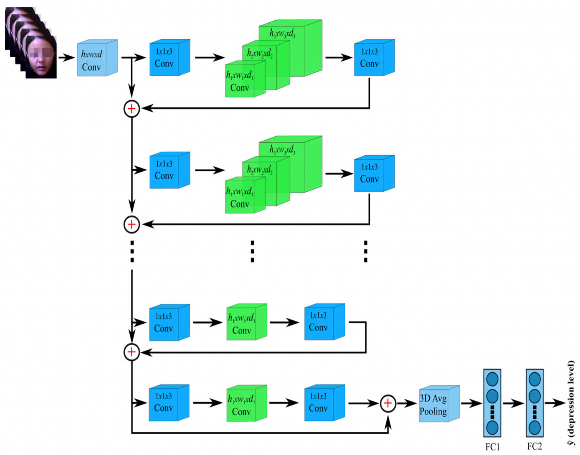


Fig 11: MSN model architecture[24]

We reimplemented the model architecture which was originally implemented in Keras using PyTorch.

We also intended to train a preliminary model to replicate the paper's results, and have contacted the AVEC workshop organizer, but have not received access to the database.

(Colab Notebook Link To The Implementation Code: https://colab.research.google.com/drive/1uQ3UWvVUFYzisLjLQ7xHB6ZhrA5y161h#scrollTo=0jWHMlw3X8LS)

6. Results and Discussion

The experiments are carried out using the Google Colab Pro + with CPU RAM of 50GB and the Nvidia P100 GPU.

6.1 Experimental results on depression classification and regression

Model	Accuracy	Recall
Custom CNN	**61%**	**84%**
Xception	**81%**	**68%**
Siamese	**69%**	**67%**
Vgg16 (audio + text)	**71%**	**69%**

Table 3: Classification results.

As we can see in Table 3, our simple 6 layer Custom CNN model performs the best when recall is considered, but the Xception model has a relatively higher accuracy when compared to the Custom CNN model. This surprising comparison between the Custom CNN and the Xception model might suggest that an architecture that has a complexity that is between these 2 models might be best suited to this problem. Furthermore, we can say that the class imbalance problem has been dealt with effectively as the model does not just predict all ‘1’s or ‘0’s. Therefore, we hypothesize that it has learnt some differentiating features between the audio spectrograms of depressed and non-depressed individuals. Although the Siamese network does not produce the best results, given the fact that it was trained only on a subset of the entire training data (due to computational constraints) shows that, if enough experimentation is done with more models and much larger training sizes, the Siamese network could be a promising solution to depression classification, especially in cases where there is limited data. Finally, we observe that the VGG 16 model which uses both audio as well as text as the input does not outperform all the other models as it did in the regression case. This suggests that further experimentation with other CNN architectures can be performed to take full advantage of this multi-modal input.

Indeed the results for audio regression with and without pre-training appears to be on par with the current depression detection research on the AVEC-2016 competition [Table 2]. When we examined the predictions, we discovered that the network can only recognize a few scores. The recognized PHQ-8 scores were in the range of 3-8, when the original score was in the range of 0-24. Because data in the 3-8 score range is abundant, the model attempts to overfit the data. However, the score range is not affected by the Audio + Text model (VGG16 + Text). For the videos, the State-of-the-art 'MAE' score is **'5.98'** and we have achieved a MAE score of 6.54. This could be improved with extra data augmentation and data preprocessing as detailed in the section below.

Model	Pretrained Model	Custom Model	Text Included	RMSE	MAE	Reason
VGG 16	**YES**	**NO**	**NO**	**76.79**	**6.74**	**Not able to generalize all PHQ-8 scores**
VGG 16	**NO**	**YES**	**NO**	**43.74**	**5.72**	**Not able to generalize all PHQ-8 scores**
VGG 16 + Text	**NO**	**YES**	**YES**	**57.65**	**6.54**	**Able to generalize all PHQ-8 scores**

Table 2: Regression Results

6.2 Analysis of Experiments:

The AVEC-2016 dataset is primarily concerned with interviewing applicants and assigning a PHQ-8 score based on their responses. 1) The audio to spectrogram conversion takes the whole duration of the audio and transforms it to a single image of fixed size, with the same spectrogram image dimension for varying audio lengths, compressing some characteristics in lengthier audio files. 2) Because the dataset contains two persons, the interviewer's tone will be neutral; nevertheless, we must assess the tone of the interviewee; therefore, representing both voices in the same spectrogram will tend towards the neutral score unless the interviewee has an extreme tone. 3) There is some noise in the audio, and despite our best efforts, the noise is still there. 4) There is a significant gap between the question and the answer; the gap is completely silent since we have encoded those features in the spectrogram images. 5) Training of the full dataset with a siamese model using various architectures such as ResNet, InceptionNet, and so on.

7. Conclusion and Future Work

We demonstrate that the Siamese network is an effective way to tackle the problem of depression classification using spectrograms. To the best of our knowledge, this approach for this problem has never been explored before.

We create a SpecAugment inspired augmentation strategy similar to the cutout data augmentation strategy, that places random masks (horizontal and vertical patches) in the frequency and time domain and show that it can result in effective and efficient data augmentation. Furthermore, it can be observed from our results that the complexity of the models we use in such cases must not be as complex as the Xception model, but at the same time not as simple as the simple Custom CNN used in our project. Instead a simple network must be taken and the complexity must be systematically increased until results start deteriorating. Finally, we show that using additional input data such as text along with audio has the potential to substantially improve our performance.

In the future, we want to create EmpathNet, an architecture that improves on the current state-of-the-art performance and propose a novel approach that combines visually interpretable representation learning and deep distribution learning, as explored in [16] and [19]. We would like to train and improve the MSN model [24] on the EDAIC dataset to analyze the correlation between facial videos and depression scores. Another idea we would like to further investigate is the use of Siamese networks for depression classification. Finally, we want to test the model with unseen data, such as videos from YouTube or movies. Intriguing research has been undertaken to investigate unique inventive concepts such as merging video input with additional subject data other than video and audio, such as electroencephalogram (EEG) data, to perform emotion recognition [26, 28]. This sparks the notion of utilizing the same multimodal data to investigate depression detection. Furthermore, research works such as [18] look at the issue of fairness in this sector, emphasizing the significance of evenly distributed datasets comprising data from different races and genders, which might be a useful contribution.

8. References

[1] James, Spencer L., Degu Abate, Kalkidan Hassen Abate, Solomon M. Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar et al. "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017." *The Lancet* 392, no. 10159 (2018): 1789-1858.

[2] Friedman, Edward S., Duncan B. Clark, and Samuel Gershon. "Stress, anxiety, and depression: Review of biological, diagnostic, and nosologic issues." Journal of anxiety disorders 6, no. 4 (1992): 337-363.

[3] Pampouchidou, Anastasia, Panagiotis G. Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Pediaditis, and Manolis Tsiknakis. "Automatic assessment of depression based on visual cues: A systematic review." *IEEE Transactions on Affective Computing* 10, no. 4 (2017): 445-470.

[4] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge," in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 3–10.

[5] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression Recognition Based on Dynamic Facial and Vocal Expression Features Using Partial Least Square Regression," in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 21–30.

[6] 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 ACM International Workshop on Au- dio/Visual Emotion Challenge*, 2013, pp. 11–20.

[7] H.P. Espinosa, H.J. Escalante, L. Villaseor-Pineda, M. Montes-y- Gmez, D. Pinto-Avedao, and V. Reyez-Meza, "Fusing Affective Dimensions and Audio-Visual Features from Segmented Video for Depression Recognition," in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 49–55.

[8] A. Jan, H. Meng, Y.F.A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic Depression Scale Prediction Using Facial Expression Dynamics and Regression," in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 73–80.

[9] H. Kaya, F. illi, and A.A. Salah, "Ensemble CCA for Continuous Emotion Prediction," in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 19–26.

[10] de Melo, Wheidima Carneiro, Eric Granger, and Abdenour Hadid. "Combining global and local convolutional 3d networks for detecting depression from facial expressions." In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1-8. IEEE, 2019.

[11] Valstar, Michel, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. "Avec 2016: Depression, mood, and emotion recognition workshop and challenge." In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pp. 3-10. 2016.

[12] R. Ranjan *et al*., "Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans," in *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 66-83, Jan. 2018, doi: 10.1109/MSP.2017.2764116.

[13] Arselan Ashraf, Teddy Surya Gunawan, Bob Subhan Riza, Edy Victor Haryanto, Zuriati Janin, "On the review of image and video-based depression detection using machine learning", Indonesian Journal of Electrical Engineering and Computer Science, vol. 19, no. 3, pp. 1677–1684, September 2020.

[14] Al Jazaery, Mohamad, and Guodong Guo. "Video-based depression level analysis by encoding deep spatiotemporal

features." IEEE Transactions on Affective Computing 12, no. 1 (2018): 262-268.

[15] Jan, Asim, Hongying Meng, Yona Falinie Binti A. Gaus, and Fan Zhang. "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions." IEEE Transactions on Cognitive and Developmental Systems 10, no. 3 (2017): 668-680.

[16] Zhou, Xiuzhuang, Kai Jin, Yuanyuan Shang, and Guodong Guo. "Visually interpretable representation learning for depression recognition from facial images." IEEE Transactions on Affective Computing 11, no. 3 (2018): 542-552

[17] Gavrilescu, Mihai, and Nicolae Vizireanu. "Predicting depression, anxiety, and stress levels from videos using the facial action coding system." *Sensors* 19, no. 17 (2019): 3693.

[18] Du, Mengnan, Fan Yang, Na Zou, and Xia Hu. "Fairness in deep learning: A computational perspective." IEEE Intelligent Systems 36, no. 4 (2020): 25-34.

[19] De Melo, Wheidima Carneiro, Eric Granger, and Abdenour Hadid. "Depression detection based on deep distribution learning." In 2019 IEEE International Conference on Image Processing (ICIP), pp. 4544-4548. IEEE, 2019.

[20] Prabhudesai, Siddharth, Apurva Mhaske, Manvi Parmar, and Sumedha Bhagwat. "Depression Detection and Analysis Using Deep Learning: Study and Comparative Analysis." In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), pp. 570-574. IEEE, 2021.

[21] He, Lang, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo et al. "Deep learning for depression recognition with audiovisual cues: A review." Information Fusion 80 (2022): 56-86.

[22] Victor, Ezekiel, Zahra M. Aghajan, Amy R. Sewart, and Ray Christian. "Detecting depression using a framework combining deep multimodal neural networks with a purpose-built automated evaluation." Psychological assessment 31, no. 8 (2019): 1019.

[23] De Melo, Wheidima Carneiro, Eric Granger, and Miguel Bordallo Lopez. "Encoding temporal information for automatic depression recognition from facial analysis." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1080-1084. IEEE, 2020.

[24] de Melo, Wheidima Carneiro, Eric Granger, and Abdenour Hadid. "A deep multiscale spatiotemporal network for assessing depression from facial dynamics." IEEE Transactions on Affective Computing (2020).

[25] Sharma, Ragini, Saman Biookaghazadeh, Baoxin Li, and Ming Zhao. "Are existing knowledge transfer techniques effective for deep learning with edge devices?." In 2018 IEEE International conference on edge computing (EDGE), pp. 42-49. IEEE, 2018.

[26] Savran, Arman, Koray Ciftci, Guillaume Chanel, Javier Mota, Luong Hong Viet, Blent Sankur, Lale Akarun, Alice Caplier, and Michele Rombaut. "Emotion detection in the loop from brain signals and facial images." In Proceedings of the eNTERFACE 2006 Workshop. 2006.

[27] Siddharth, Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski. "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing." IEEE Transactions on Affective Computing (2019).

[28] Ma, Xingchen, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. "Depaudionet: An efficient deep model for audio based depression classification." In Proceedings of the 6th international workshop on audio/visual emotion challenge, pp. 35-42. 2016.

[29] Chlasta, Karol, Krzysztof Wołk, and Izabela Krejtz. "Automated speech-based screening of depression using deep convolutional neural networks." Procedia Computer Science 164 (2019): 618-628.

[30] Rodrigues Makiuchi, Mariana, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. "Multimodal fusion of BERT-CNN and gated CNN representations for depression detection." In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, pp. 55-63. 2019.

[31] Srimadhur, N. S., and S. Lalitha. "An end-to-end model for detection and assessment of depression levels using speech." Procedia Computer Science 171 (2020): 12-21.

[32] Wu, Yue, and Qiang Ji. "Facial landmark detection: A literature survey." International Journal of Computer Vision 127, no. 2 (2019): 115-142.

[33] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[34] Koch, Gregory, Richard Zemel, and Ruslan Salakhutdinov. "Siamese neural networks for one-shot image recognition." In ICML deep learning workshop, vol. 2, p. 0. 2015.

[35] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[36] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

[37] Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching word vectors with subword information." Transactions of the association for computational linguistics 5 (2017): 135-146.

[38] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[39] Kilshore Prahallad, “Spectrogram, Cepstrum and MelFrequency Analysis,” Carnegie Mellon University.

[40] Rao, K. Sreenivasa, and K. E. Manjunath. Speech recognition using articulatory and excitation source features. Springer, 2017.

[41] Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D, Marsella S, Traum DR. The Distress Analysis Interview Corpus of Human and Computer Interviews. In LREC 2014 May (pp. 3123-3128) Ringeval, Fabien, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt et al. "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition." In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, pp. 3-12. ACM, 2019.