# Machine Learning-enabled Depression Detection Through Audio, Visual and Textual Data

Prasanth Bathala, Janavi Khochare, Nikhil Viswanath Sivakumar
*School of Electrical and Computer Engineering*
*Georgia Institute of Technology*
Atlanta, GA, USA
pbathala3@gatech.edu, janavikhochare@gatech.edu, nsivakumar33@gatech.edu

***Abstract*—Traditionally, depression was identified through in-depth clinical interviews, during which the psychologist would analyze the subject's responses to ascertain his or her mental condition. By combining the three modalities of word context, audio, and video in our model, we try to imitate this strategy and predict an output relating to the patient's mental state. To account for the subject's level of depression, the output has been separated into many levels. We have developed a deep learning model that combines all three modalities, gives them the proper weights, and produces an output. The objective of this work is to apply multiple machine-learning models, including Support Vector Machine (SVM) [1], Convolutional Neural Network (CNN) [2], and Long Short-Term Memory (LSTM) [3], to both individual and combination modalities and do a comparative analysis. A comparison of the model performances shows that applying hybrid fusion performs better than training on individual modalities, and the gating mechanism in LSTM improves results. In particular, LSTM with sentence-level gating outperforms the other algorithms by a sizeable margin, with a test accuracy of 63 percent.**

***Index Terms*—depression detection, CNN, LSTM, gating mechanism, hybrid fusion**

## I. Introduction

An effective, self-contained, and easily accessible method of identifying depression is the need of the hour. A growing number of people are becoming depressed as society shifts toward increasingly stressful environments. Only by detecting it will we be able to work on curing it. Our primary motivation is to develop such a model. Clinical interviews with the individuals are required to generate the three modalities to evaluate our model (as input to our model). Extensive research in this sector has demonstrated that a sad patient exhibits several nuanced indicators, which can be detected more accurately by investigating all three modalities at the same time. Numerous physiological and metabolic changes might result from a change in mental behavior. According to research, people who are depressed tend to stutter when they speak, causing unnatural pauses to appear in their speech. Another feature that the topic emphasizes is more instances of erroneous pronunciation. Video technology can be used to identify additional characteristics, such as unusual eye contact, less frequent lip movements, altered posture, and so on. Lexical analysis can be used to examine the subject's speech in context, which also reveals crucial details about his or her mental state. As a result, a more general model that takes into account all of these elements can be developed by merging all of these channels. As a result, due to the availability of more reliable components, better forecasts can be made. Several difficulties that can be anticipated from this model include the following:

1) Since our model is essentially a DL model, a lot of data is needed in all three modalities.
2) Aligning these three modalities based on their timestamps presents another difficulty. To comprehend the association between various modalities, it is crucial that our model receive them simultaneously.
3) Since video processing is involved, a lot of computing power will be required to train our model.

A late fusion detection approach is constructed for model prediction in D. Huang's regression method, which is based on PLS [4]. A multimodal HCRF model by D. Devault that utilizes question-answer pairs has been developed. They are analyzed for model prediction [5]. The same method is used by Gong and others. On top of that, he incorporates his multimodal method with the question-answer-based model, taking into account all three modalities for model prediction [6]. Sun et al. carry out related research as well. They developed a single-model random forest-based classifier that employs a question-and-answer methodology. Model prediction is done using this classifier [7]. Ma et al. suggest using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for a higher-level audio representation in their audio-based depression classification technique. Ma et al. only utilize audio as their modality. Their research utilizes a CNN to input the audio-based data, and then an LSTM network for model prediction [8].

The current approach [9] for this problem investigates a model-based optimal fusion, i.e, it focuses more on how much each modality should influence the outcome rather than employing early or late fusion techniques. After extraction, early fusion essentially combines feature vectors from each modality into a single vector and feeds it to the model to learn outcomes. In late fusion method, we train separate models for each modality and subsequently merge their output by assigning weights to each model. The fact that learned representation of 1 modality might be damaged by the other modality is something that both of these theories overlook.

## II. METHODOLOGY

The flow and system analysis is shown in Fig. 1.

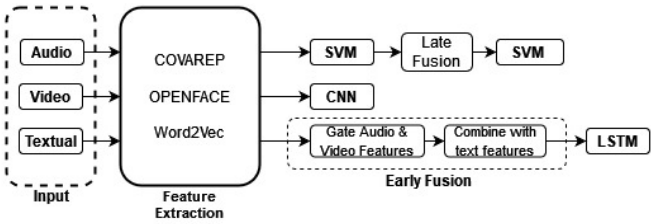


Fig. 1. Flowchart for depression detection system

### A. Data Processing

The deployed dataset, DAIC-WOZ, is a subset of the bigger corpus, the Distress Analysis Interview Corpus (DAIC) [10]. It includes clinical interviews intended to help the treatment of psychological distress illnesses such as anxiety, depression, and post-traumatic stress disorder. The data collected includes thorough questionnaire responses as well as audio and video recordings. The Wizard-of-Oz interviews are also part of the DAIC-WOZ dataset. They were performed by Ellie, an animated virtual assistant that was operated from another room by a real interviewer. For a number of verbal and non-verbal aspects, the data has been transcribed and annotated. An interaction transcript, participant audio files, and facial feature extraction from the recorded video are all included in each participant's session. The dataset contains 189 sessions of interactions, ranging anywhere from 7 to 33 minutes. It contains interviews with 59 depressed and 139 non-depressed subjects.

*1) **Textual Data**:* The entire interview conversation with the patient is transcripted and available in the form of a CSV file. Sentences have been timestamped and further categorized according to the speaker. Speech overlap is indicated by overlapping timestamps. If speaking is interrupted, the entire word that was planned to be said is written, followed by a note that includes the part that was actually spoken, enclosed in angle brackets: people. <peop>. The comment is only intended for human readers; the entire word transcription was used to prevent confusing the models by training them on non-words.

*2) **Audio Data**:* The Audio features are extracted using the COVAREP toolbox [11]. All audio elements, including the formants, are played every 10 ms. As a result, the audio features are sampled at 100 Hz. 12 Mel-frequency cepstral coefficients (MFCCs)—F0, VUV, NAQ, QOQ, H1H2, PSP, MDQ, peak slope, Rd, Rdonf, MCEP0-24, HMPDM0-24, and HMPDD0-12—are among the features. Pitch tracking, peak slope, maximal dispersion quotients, and glottal source characteristics are supplied in addition to the MFCCs. The current sample's voiced or unvoiced status is indicated by the feature flag VUV (voiced/unvoiced). Features including F0, NAQ, QOQ, H1H2, PSP, MDQ, peak slope, and Rd are zeroed out for an unvoiced sample.

*3) **Video Data**:*
The face features from the participant interview videos were retrieved using OpenFace [12] and are included in the dataset. The facial features consist of 68 2D points on the face, 24 AU features that assess facial activity, 68 3D points on the face, 16 features to represent the subject's gaze, and 10 features to describe the subject's stance. A total of 388 video features are contained in this.

### B. Models

The dataset is skewed with a 7:3 ratio, of non-depressed class to depressed. To increase the bias of the model, we used bootstrap sampling on the data.

*1) **Support Vector Machine**:* Support Vector Machine (SVM) is used to classify whether a given sample is depressed or not. For each participant, we have nearly 40,000 to 50,000 audio and video samples. We averaged all samples for each session. For the text data, we translated each word into 300 vector-size word embeddings using word2vec [13]. Further, the 3D data (sentences x words x 300 features) for each session were first averaged over each word and then flattened. The processed data is then used for training the SVM with an RBF Kernel. We trained SVM on three modalities separately and then the decision labels from each modality are trained on another SVM model to perform late fusion.

*2) **Convolutional Neural Networks**:* A Convolutional Neural Network (CNN) model comprising 6 layers was built with the first 4 layers of conv2D for text modality and conv1D layers for audio and video modality and a max pooling layer. Further, flattening and fully connected layers were added with the ReLU activation function. Sigmoid activation was used in the last layer. For audio and video modality, the first 40,000 samples are taken for each session respectively. These values are chosen based on computational availability. For the textual data, the resulting word embeddings from word2vec are sampled based on the threshold set for the maximum number of words and sentences.

*3) **LSTM model with/without gating (Sentence Level)**:*
We carried out an early fusion [14] for the Long short-term memory (LSTM) model by initially selecting sentences from the transcripts. The audio and samples are then averaged over timestamps for each sentence. This process is called Sentence level force alignment. The audio and visual features were then gated using highway layers. Each highway layer consists of two non-linear transforms: a carry and a transform gate that specify how much of the output is created by modifying the input and how much information should advance. The matching text features are linked after the audio and video data have been separately sent to the sentence-to-highway layers.

The final output is obtained by passing the concatenated vector through an LSTM.

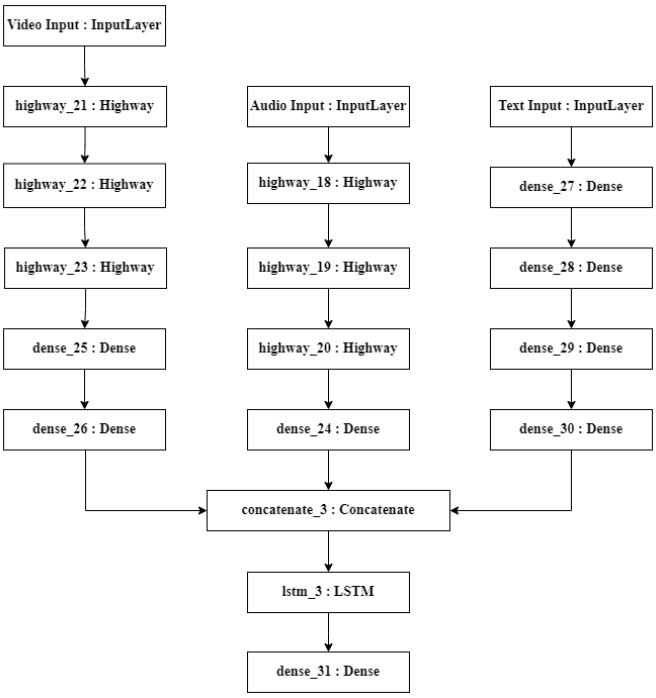


Fig. 2. Model Architecture

The LSTM's fundamental model architecture is depicted in Fig. 2. The three feedforward highway layers are first traversed by audio and visual features. The dimensionality of both audio and video features is then reduced using dense layers. LSTM with 128 hidden nodes is employed after concatenation with the appropriate text features. The output is then obtained by applying a dense layer with a sigmoid activation function. The learning rate was 0.0001. The model is terminated using the early stopping callback in Keras API.

*4) **LSTM with Gating (Word Level)**:* Data was forcibly aligned in this case at the word level, and trained in the LSTM model. This procedure is called Word-level force alignment. According to the number of words and character length present in the sentence, we employed a unitary approach to convert the sentence-level time stamps to word-level time stamps. Highway layers were then used to gate the audio and visual characteristics. A carry and a transform gate, which determine how much of the output is formed by changing the input and how much information should advance, are the two non-linear transformations that make up each highway layer. After the audio and video data have been delivered individually to the sentence-to-highway layers, the corresponding text features are linked. The vector that has been concatenated is placed through an LSTM to produce the result.

The model architecture for the word model is similar to the sentence as shown in Fig. 2. The model architecture for the word-level multi-modal fusion is shown in Fig. 3. $D_t^v$ and $D_t^a$

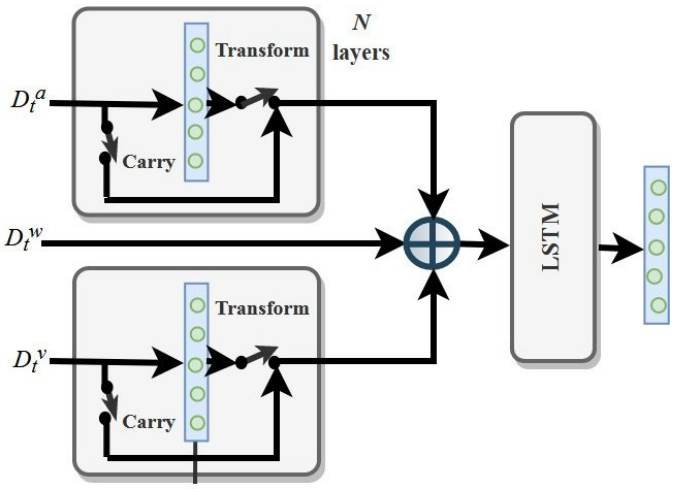


Fig. 3. Word-level Multimodal Fusion with Gating [15]

are the audio and video features, $D_t^w$ is the textual (words) data in Fig. 3. First, audio $(D_t^a)$ and visual $(D_t^v)$ characteristics travel across the three feedforward highway layers. Then, utilizing dense layers, the dimensionality of both audio and video information is decreased. Following concatenation with the necessary text (words) features $(D_t^w)$, an LSTM with 128 hidden nodes is used. The application of a dense layer with a sigmoid activation function results in the output. The learning rate was 0.0001. Using the early stopping callback in the Keras API, the model is terminated.

## III. RESULTS

A sum total of 6 different models are run for the 3 modalities (text, audio, and video). SVMs and CNNs are trained initially on individual modalities. For the cases of fusion (for SVM), the modalities are combined and given as training data. LSTM models are trained on two types of varieties - with and without gating, and for sentence-level and word-level features.
Each model is considered a binary classification task, where class 0 corresponds to 'not depressed' and class 1 corresponds to 'depressed'. The performance metrics for each model, with fusion and gating mechanisms, are shown in Table I. The metrics chosen to measure and compare the performances of the different models were accuracy, precision, and F1-score. The formulae of these metrics are given in Eqns. 1-3:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{1}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{2}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

| Model | Modality | Accuracy | Precision | F1-Score |
|---|---|---|---|---|
| SVM | Text | 0.28 | 0.37 | 0.24 |
| | Audio | 0.39 | 0.45 | 0.41 |
| | Video | 0.35 | 0.45 | 0.58 |
| SVM with Late Fusion | Text + Audio + Video | 0.45 | 0.47 | 0.60 |
| CNN | Text | 0.44 | 0.47 | 0.46 |
| | Audio | 0.47 | 0.46 | 0.51 |
| | Video | 0.30 | 0.42 | 0.50 |
| LSTM without Gating for Sentence-level | Text + Audio | 0.35 | 0.42 | 0.36 |
| | Text + Video | 0.63 | 0.48 | 0.60 |
| | Text + Audio + Video | 0.59 | 0.45 | 0.57 |
| LSTM with Gating for Sentence-level | Text + Audio | **0.63** | 0.36 | **0.61** |
| | Text + Video | **0.63** | 0.60 | **0.61** |
| | Text + Audio + Video | **0.59** | 0.57 | **0.59** |
| LSTM with Gating for Word-Level | Text + Audio + Video | 0.58 | 0.45 | 0.56 |

TABLE I: Performance Metrics for Various Models

- **SVM:** SVMs are observed to provide the least impressive prediction. The poorer performance of this machine learning model could be attributed to the fact that the SVM models used perform an averaging operation over the modalities. This could lead to a possible loss of information, and hence affects the model learning.

- **SVM with late fusion:** Upon adding the fusion method to SVMs, we observe this model provides much better results than training SVMs on individual modalities. This is because the fusion method combines the learnt output decisions of each modality (ran in individual SVM models) and trains these scores in a new SVM model. This essentially means that the errors from multiple models are dealt with independently, ensuring the overall errors are uncorrelated.

- **CNN:** CNN is generally noticed to perform better than the individual modalities trained on SVMs. Being a deep learning model, CNNs learn the temporal data present in the text, audio, and video modalities. Additionally, CNNs were created to train on 2D and 3D image data. Hence, in our case, CNNs provide better results for audio and video data. CNN model over combined modalities with fusion is expected to exceed the results of individual modalities trained over SVM, fusion over SVM and individual modalities trained over CNN.

- **LSTM with and without gating for Sentence-level features:** The LSTM models perform better than both SVMs and CNNs. LSTMs are a variation of Recurrent Neural Networks, which are used in literature to learn sequential data well. LSTMs in particular, with their gating architecture, help in memorizing patters. This is especially key in the case of textual and audio data, thereby helping in modeling and understanding patterns in the sequential data in this case. This could be due to the fact that the gating mechanism adds weights to the features. Hence, this feature extraction technique enhances the contribution of important features and nullifies those features which do not lead to better prediction. Therefore, we notice that LSTM with gating features performs better in the case of sentence-level features.

- **LSTM with gating for Word-level features:** Additionally, word-level features were experimented with LSTM models to observe for any improvement in detecting depression. Adhering to the discussions of the gating mechanism over sentence-level features, it is seen that gating with word-level features provides healthy results. Though the prediction rate is decent, it does not exceed the prediction of models trained over sentence-level features. The reason for this is, word-level features do not utilize and understand the entire context of the conversation between the patient and the interviewer which is recorded as the data. Whereas in the case of sentence-level features, the entire context of the data is understood.

## IV. Conclusion

A system of models comprising various machine learning models was designed to detect depression in a patient by training the models on the person's textual, audio, and video data modalities. Through various data processing techniques to utilize textual features, and learning audio and visual features through deep and complex models, a binary classification of detecting a case of depression or not was performed. Additionally, the inclusion of fusion techniques and gating mechanisms were used over the initial models, which were also trained over both sentence-level and word-level features. It was concluded that out of all the models shown in this paper, the long-short term memory models using a gating mechanism over sentence-level features provide the best detection of depression given the three modalities. This study helps towards the right treatment of depressed people, with the necessity to create awareness in

the study of mental health as well as advertise the importance of mental health among researchers. Future work in this field could be done towards more processing techniques and more variations over the deep learning models mentioned in the paper as well as in other complex models available in the literature.

## REFERENCES


[1] Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their applications 13.4 (1998): 18-28.

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Communications of the ACM 60.6 (2017): 84-90.

[3] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[4] Meng, Hongying, Di Huang, Heng Wang, Hongyu Yang, Mohammed Ai-Shuraifi, and Yunhong Wang. "Depression recognition based on dynamic facial and vocal expression features using partial least square regression." In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, pp. 21-30. 2013.

[5] Yu, Z., Scherer, S., Devault, D., Gratch, J., Stratou, G., Morency, L. P., & Cassell, J. (2013). Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs. In Semdial 2013 DialDam: Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue (pp. 160-169).

[6] Gong, Yuan, and Christian Poellabauer. "Topic modeling based multimodal depression detection." Proceedings of the 7th annual workshop on Audio/Visual emotion challenge. 2017.

[7] Sun, Bo, Yinghui Zhang, Jun He, Lejun Yu, Qihua Xu, Dongliang Li, and Zhaoying Wang. "A random forest regression method with selected-text feature for depression assessment." In Proceedings of the 7th annual workshop on Audio/Visual emotion challenge, pp. 61-68. 2017.

[8] 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.

[9] Rohanian, Morteza, Julian Hough, and Matthew Purver. "Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs." arXiv preprint arXiv:2106.15684 (2021).

[10] Gratch, Jonathan, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood et al. The distress analysis interview corpus of human and computer interviews. UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES, 2014.

[11] Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S. (2014, May). COVAREP—A collaborative voice analysis repository for speech technologies. In 2014 ieee international conference on acoustics, speech and signal processing (icassp) (pp. 960-964). IEEE.

[12] Baltrušaitis, Tadas, Peter Robinson, and Louis-Philippe Morency. "Openface: an open source facial behavior analysis toolkit." 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016.

[13] Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[14] Boulahia, Said Yacine, Abdenour Amamra, Mohamed Ridha Madi, and Said Daikh. "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition." Machine Vision and Applications 32, no. 6 (2021): 1-18.

[15] Rohanian, Morteza, Julian Hough, and Matthew Purver. "Detecting Depression with Word-Level Multimodal Fusion." INTERSPEECH. 2019.


## V. SPECIFIC CONTRIBUTION

- **Janavi Khochare:** Worked on data processing for textual features and trained them over SVM, CNN, and LSTM models and analysed combined modalities for SVM with late fusion.
- **Prasanth Bathala:** Worked on audio feature extraction and trained over SVM, CNN, and LSTM models; compared sentence-level and word-level features.
- **Nikhil Viswanath Sivakumar:** Worked on video features for SVM, CNN, and LSTM models and analyzed the various models along with gating mechanism