

A
Project Report
On
Depression Detection Using Multiple Modalities

(CE356 - Software Group Project-III)

Prepared by
21CE001 – Harshit Rajnikant Ajakiya
21CE002-Andrew Babu Augustine

Under the Supervision of
Prof. Aayushi Chaudhary

Submitted to
Charotar University of Science & Technology (CHARUSAT)
for the Partial Fulfillment of the Requirements for the
Degree of Bachelor of Technology (B.Tech.)
in U & P U. Patel Department of Computer Engineering (CE)
for B.Tech Semester 5

Submitted at



Accredited with Grade A+ by NAAC



U & P U. PATEL DEPARTMENT OF COMPUTER ENGINEERING
Chandubhai S. Patel Institute of Technology (CSPIT)
Faculty of Technology & Engineering (FTE), CHARUSAT
At: Changa, Dist: Anand, Pin: 388421.
October, 2023



CHARUSAT
CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

Accredited with Grade A+ by NAAC

CERTIFICATE

This is to certify that the report entitled Depression Detection using Mutli-Modalities is a bonafied work carried out by **21CE002,21CE001** under the guidance and supervision of **Prof. Aayushi Chaudhary** for the subject **Software Group Project-III(CE356)** of 5th Semester of Bachelor of Technology in **Computer Engineering** at Chandubhai S. Patel Institute of Technology (CSPIT), Faculty of Technology & Engineering (FTE) – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate themself have duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred by the examiner(s).

Under the supervision of,

Prof.Aayushi Chaudhary
U & P U. Patel Dept. of Computer Engineering
CSPIT, FTE, CHARUSAT, Changa, Gujarat

Dr. Nikita Bhatt
Head - U & P U. Patel Department of Computer Engineering,
CSPIT, FTE, CHARUSAT, Changa, Gujarat.

Chandubhai S. Patel Institute of Technology (CSPIT)
Faculty of Technology & Engineering (FTE), CHARUSAT

At: Changa, Ta. Petlad, Dist. Anand, Pin: 388421. Gujarat

Abstract

Conventionally, depression detection has relied on extensive clinical interviews, where psychologists analyse the subject's responses to determine their mental state. In our model, we aim to integrate this approach by combining three modalities: word context, audio, and video, to predict the patient's mental health status. The model's output is categorized into different levels to assess the severity of the subject's depression.

We aim at developing a deep learning model that fuses these three modalities, assigning them appropriate weights, resulting in an output. This fusion approach addresses the following key issues:

1. **Presence of Noise in Modalities:** Our model can handle the presence of noise in any of the modalities by jointly considering multiple sources of information.
2. **Control over Modality Contribution:** We can control the level of contribution from each modality, allowing us to emphasize the most informative aspects of the data for more accurate predictions.

This fusion-based approach enhances the robustness and effectiveness of depression detection by leveraging the complementary information from multiple modalities.

Table of Contents

Abstract.....	i
List of Figures	iii
Chapter 1 Introduction.....	1
Chapter 2 Literature Review	2
2.1 PHQ8	4
2.2 What is modality	5
2.3 What is late fusion	6
Chapter 3 Dataset	7
3.1 DAIC-WOZ DATSET	7
3.2 Modalities	7
3.2.1 Video Modality	7
3.2.2 Audio Modality	7
3.2.2 Text Modality.....	7
3.3 Dataset Size	7
3.4 Explanation of Terms and Files.....	8
Chapter 4 Proposed solution.....	10
Chapter 5 Work Done and Results.....	12
5.1 CNN.....	12
References.....	14

List of Figures

Fig 3.4.1 File Structure with Patient wise folder	22
Fig 3.4.3 Individual files under each Patient	22

Chapter 1: Introduction

An accurate, autonomous, accessible approach to detect depression is the need of the hour. As society moves to more and more stressful environments, a higher percentage of the population is developing depressing tendencies. Only if we're able to detect it, can we work to cure it. The motivation to create such a model is our driving factor.

To test our model, clinical interviews of the subjects need to be done for generation of the 3 modalities (as input to our model). It has been noted, through extensive research in this field, that a depressed subject displays various intricate signs, that can be better caught by studying all the 3 modalities together. Due to a change in mental behaviour, various physiological and physiological changes can be detected. Research shows that a depressed subject often stammers while talking and thus uneven pauses can be caught in their speech. More occurrences of incorrect pronunciation are another attribute that the subject showcases. Using video modality, other factors such as abnormal eye contact, less frequent mouth movement, changed posture, etc can be caught. Using lexical analysis, the context of the words spoken by the subject can be analysed, which also provides essential information regarding his/her mental health. Thus, integrating all these channels, a more generic model can be built, which takes all these factors into consideration. Thus, better predictions can be made due to the presence of more viable factors.

Certain challenges that can be expected out of this model are:

- As our model is basically a DL model, large amount of dataset is required, in all 3 modalities.
- Alignment of these 3 modalities, according to their timestamp is another challenge. It's of utmost importance that our model receives these modalities in sync, to understand the correlation between them.
- Since video processing is involved, thus large amounts of computation power will be needed to train our model.

Chapter 2: Literature Review

D. Huang uses a regression method based on PLS wherein a late fusion detection method is built for model prediction[1].

D. Devault has built a multimodal HCRF model which works on question-answer pairs. It analyzes them for model prediction[2].

Gong et. al. use the same approach. Building on it, he combines the question-answer based model with his multi-modal approach, taking into consideration all the 3 modalities for model prediction[3].

Similar work is also done by Sun et al. They built a single model random forest-based classifier which works on the question-answer based approach. This classifier is used for model prediction[4].

Ma et al. propose an audio-based method for depression classification using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for a higher-level audio representation. Ma et al. works only on the audio based modality. He inputs the audio based data into a CNN and then further uses an LSTM network for model prediction[5].

In the work done by Shivakumar et al. [6], the temporal nature of audio/visual modalities is considered using a window-based representation instead of frame-level analysis.

Utilizing complementary information from the text and audio features, J. Glass et al. proposed a model in which different LSTM branches for each modality are integrated via a feed-forward network [7]. However, while this work tries to predict depression based on late or early fusion methods [1, 3] or the sequential nature of their inputs [6, 7], learning the time-dependent relationships between language, visual and audio features in detecting depression is still unexplored.

The major problems that these approaches face are the different predictive power of modalities and types of noise in the representation. In previous works, gating mechanism has shown to be effective in determining the predictive power of each modality.

Another approach[10] for the same problem explores paralinguistic, head pose, and eye gaze behaviors. During the research phase, the authors found out that there are many physical attributes/changes that can be detected through appropriate sensors when a subject is depressed. In this model, the authors try to detect features like dropped speech rate, lower articulation rate (speech features), lesser eye contact, wavering eyes (eye features), and bent head (head features).

To detect speech-related features, the extraction from the dataset was done using a 2-fold approach, i.e., Manual Labelling and Automated labeling (using manually segmented data). A total of 63 statistical features were extracted through manual labeling, and 19 speaking rate features were extracted using automated labeling (using PRAAT). For eye features, it was done by training a specialized CV application that is able to detect different attributes of the eye such as the eyelids, pupil, and its extremities.

Using this, it is able to make mathematical calculations that lead to features such as blink time, blink frequency, gaze direction (left-right, up-down), etc. Previously, it was found that slower and less frequent head movements, increased eye contact avoidance, and less social engagement with the clinical examiner are likely to also show in other social interactions.

To extract head pose and movement behavior, the face had to be detected and tracked before a 3 degrees of freedom (DOF) head pose could be calculated (yaw, roll, and pitch). A subject-specific face active appearance model was trained and built, where 30 images per subject were selected for manual annotation, then used for the face model.

These 3-DOF pose features, as well as their velocity and acceleration, were extracted to give a total of nine low-level features per frame. All of these eye and head duration features were detected when the feature in question is higher or lower than the average of the feature in question plus or minus the standard deviation of that feature for each subject's interview.

For the method in this paper, the base of the model is an SVM classifier. It is used to classify the features into binary classes, i.e., Yes (Depressed) or No (Not depressed). The extracted features are further sifted using feature extraction/Dimensionality reduction techniques like Statistical Analysis using the t-test algorithm and Principal component analysis.

Every feature was also normalized to bring down to one scale. For fusion, early, late, and hybrid fusions are explored in this paper. For early fusion, feature fusion is explored, which is basically concatenating extracted features

from the raw data. In late fusion, results from each modality are combined after training them separately. This was done on labels (decision fusion) and scores (score fusion) from the classifier. In this paper, a comparatively new fusion technique is also explored, which is hybrid fusion. In hybrid fusion, feature fusion of all modalities is performed first to create a new modality, which is then treated as an additional individual modality. The scores/decisions of this new modality are then fused with the scores/decisions of the individual modalities in either one or two levels. The dataset taken in this paper was relatively small due to which the results weren't conclusive.

The most recent approach [8] for this problem explores a model-based optimal fusion. Instead of using early fusion or late fusion technique, it focuses more on how much each modality should have an impact on the final result. Early fusion is basically concatenating the feature vectors of each modality after extraction into a single vector and feeding them to the model to learn the results. In the late fusion technique, we train individual models for each modality and then combine their results to get a final output by giving them some weights. What both of these approaches ignore is that learned representation of one modality can be undermined by the other modalities.

2.1 PHQ8 - Patient Health Questionnaire-8

The PHQ8 stands for Patient Health Questionnaire-8. It is a self-assessment tool used to screen for depression in adults. The PHQ8 is a brief questionnaire that asks about 8 common symptoms of depression, such as:

- Feeling down, depressed, or hopeless
- Loss of interest or pleasure in activities once enjoyed
- Changes in appetite (weight loss or gain unrelated to dieting)
- Trouble sleeping (either too much or too little)
- Feeling slowed down or restless
- Fatigue or loss of energy
- Feeling worthless or guilty
- Trouble thinking, concentrating, or making decisions

The PHQ8 is scored on a scale of 0 to 24, with higher scores indicating more severe depression. A score of 10 or more is considered to be a positive screen for depression, meaning that the person is likely to have depression.

The PHQ8 is a valid and reliable tool for screening for depression. It is easy to use and can be completed in just a few minutes. The PHQ8 can be used by healthcare providers, mental health professionals, and even by individuals themselves to screen for depression.

If you score 10 or more on the PHQ8, it is important to talk to your healthcare provider about your symptoms. Depression is a treatable medical illness, and there are many effective treatments available.

Here is a table that shows the scoring instructions for the PHQ8:

Score	Depression Severity
0-2	Normal
3-5	Mild
6-8	Moderate
9-12	Severe

2.2 What is a Modality in Machine Learning:

In the context of machine learning, the term "modality" refers to the different types or forms of data that can be used as input for a model. It represents the representation or format of the data being analyzed.

Data can be categorized into different modalities based on its nature and characteristics. Some common modalities in machine learning include:

- Numerical Modality: Data consisting of numerical values (e.g., age, temperature).
- Categorical Modality: Data with distinct categories (e.g., gender, color).
- Text Modality: Natural language text (e.g., sentences, documents).
- Image Modality: Visual representations (e.g., photographs).
- Audio Modality: Sound signals or recordings.
- Time Series Modality: Data with chronological order (e.g., stock prices).
- Graph Modality: Data representing relationships between entities (e.g., social networks).

Each modality may require specific processing techniques and models tailored to its characteristics.

2.3 What is Late Fusion in ML:

Late fusion, also known as late integration, is a technique used in machine learning to combine multiple modalities or sources of information at a later stage of the model pipeline. It involves merging the outputs or representations obtained from different models or modalities to make a final decision or prediction.

In late fusion, each modality or source of information is processed independently using separate models or algorithms. After processing each modality, the outputs or representations are combined or fused together to obtain a joint representation. The fusion can be performed using various techniques, such as concatenation, averaging, weighted sum, or more complex methods like attention mechanisms or neural network architectures.

Late fusion is flexible and allows each modality to be processed with specialized models, capturing unique information effectively. It is commonly used in multi-modal tasks where multiple sources of information need to be integrated, improving overall model performance and robustness. Other fusion techniques like early fusion or hybrid fusion may also be suitable depending on the task and data characteristics.

Chapter 3: Dataset

3.1. DAIC-WOZ DATASET

The DAIC-WOZ dataset [9] was collected by the University Of Southern California. It is a part of a larger DAIC (Distress Analysis Interview Corpus) that contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and PTSD.

3.2. Modalities

The dataset contains audio and video recordings and extensive questionnaire responses. Additionally, the DAIC-WOZ dataset includes the Wizard-Of-Oz interviews, conducted by an animated virtual assistant called Ellie, who is controlled by a human interviewer in another room. The data has been transcribed and annotated for a variety of verbal and non-verbal features. Each participant's session includes a transcription of interaction, participant audio files, and facial features extracted from the recorded video.

3.2.1 Video Modality

The dataset contained facial features from the videos of the participant. The facial features consisted of 68 2D points on the face, 24 AU features that measure facial activity, 68 3D points on the face, 16 features to represent the subject's gaze, and 10 features to represent the subject's pose. This made for a total of 388 video features.

3.2.2 Audio Modality

The audio features are for every 10ms; thus, the features are sampled at 100Hz. The features include 12 Mel-frequency cepstral coefficients (MFCCs), these are F0, VUV, NAQ, QOQ, H1H2, PSP, MDQ, peakSlope, Rd, Rdconf, MCEP0-24, HMPDM0-24, HMPDD0-12. Along with the MFCCs we also have features for pitch tracking, peak slope, maximal dispersion quotients, glottal source parameters. Additionally, the VUV (voiced/unvoiced) feature flags whether the current sample is voice or unvoiced. In the case where the sample is unvoiced (VUV = 0), F0, NAQ, QOQ, H1H2, PSP, MDQ, peakSlope, and Rd are set to 0.

3.2.3 Text Modality

The textual modality contains the transcript for the whole conversation of the patient with the RA in CSV format. Individual sentences have been timestamped and further classified on the basis of their speaker. Expressions like laughter, frown, etc., have been added in angular brackets as and when they occur (for example, [Laughter]). Differentiation between long/short pauses has not been made. Only word (not phoneme) level segmentation has been recorded.

3.3. Dataset size

The dataset contains 189 sessions of interactions, ranging anywhere from 7 to 33 minutes. The dataset contains interviews with 59 depressed and 130 non-depressed subjects.

3.4. Explanation of Terms and Files

CLNF_AU (Constrained Local Neural Fields Action Units):

- **Concept:** Action Units (AUs) are a set of facial muscle movements used to describe facial expressions. The CLNF_AU file likely contains data related to the presence or intensity of specific facial AUs for each patient in the dataset.
- **Usage:** This data can be crucial for analysing the emotional state of individuals, as different AUs are associated with different emotions. Machine learning models can use this data to identify patterns and make predictions related to depression based on facial expressions.

CLNF_Features (Constrained Local Neural Fields Features):

- **Concept:** CLNF features usually refer to facial landmark points and their coordinates on the face. These landmarks are used to describe the shape and position of various facial features.
- **Usage:** These features are valuable for tasks like facial expression recognition and emotion analysis. Machine learning models can use this information to understand and interpret the geometry of the face, which can be indicative of emotional states in the context of depression detection.

CLNF_Features3D (Constrained Local Neural Fields 3D Features):

- **Concept:** CLNF 3D features are likely an extension of the previous CLNF features, but in this case, they provide three-dimensional coordinates of facial landmarks.
- **Usage:** 3D facial landmarks can offer more information than 2D landmarks and are particularly useful for applications like head pose estimation and tracking facial movements, which can be relevant to assessing a person's emotional state.

CLNF_Gaze (Constrained Local Neural Fields Gaze):

- **Concept:** Gaze information typically refers to the direction in which a person is looking. It can be determined by the positions of the eyes, especially the pupils.
- **Usage:** Gaze data can be important for understanding a person's attention and engagement. For depression detection, it might be used to analyse whether individuals show patterns of reduced engagement or interest in their surroundings.

CLNF_Pose (Constrained Local Neural Fields Pose):

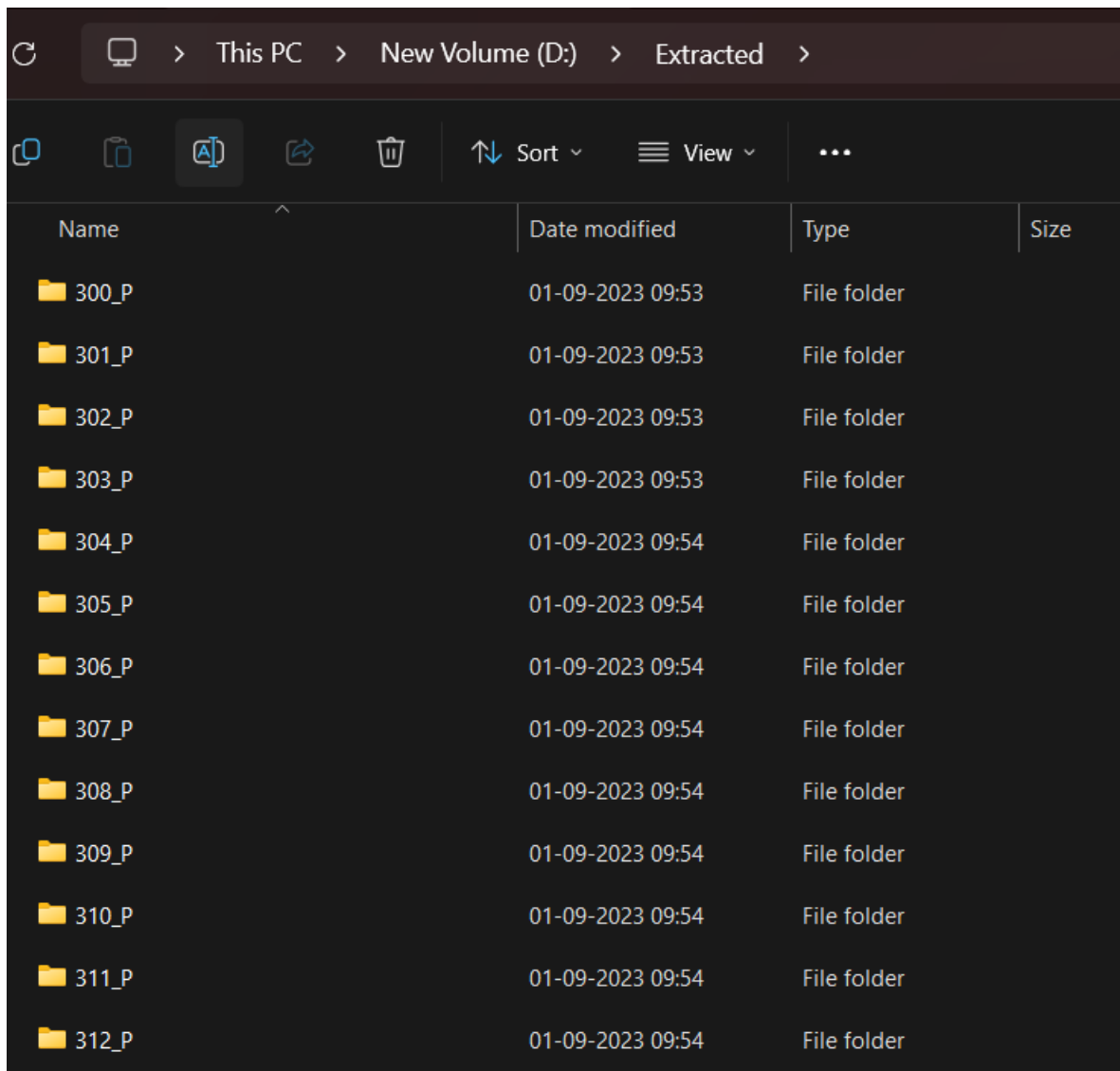
- **Concept:** Pose refers to the orientation and position of the head or face in space. It can involve information about head tilt, rotation, and position.
- **Usage:** Pose data can help assess non-verbal cues related to depression, such as a downward head tilt or lack of face-to-face interaction. Machine learning models can use this information to detect such cues and make predictions.

COVAREP (COrpus VALEncia REsearch on PErception and Design):

- **Concept:** COVAREP is a set of acoustic features typically used in speech processing and analysis. These features might include pitch, intensity, and various spectral features.

- **Usage:** While not directly related to facial data, COVAREP features can be used to analyze aspects of speech and voice that are indicative of depression. Changes in pitch, tone, or speech patterns can be associated with emotional states.

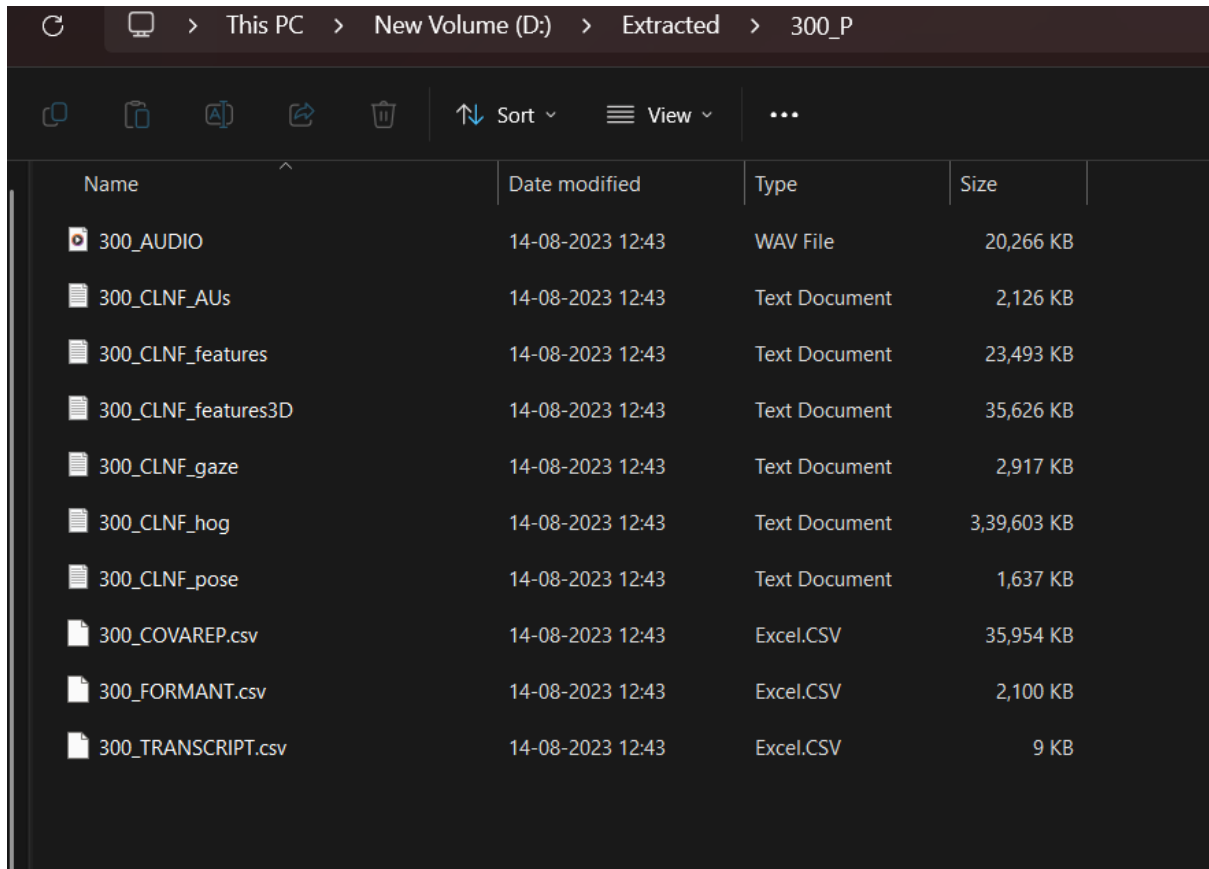
In a depression detection context, these files and concepts collectively provide a rich source of data for machine learning models to analyze and detect potential signs of depression based on facial expressions, gaze, pose, and speech features. The combination of these modalities can contribute to a more comprehensive assessment of an individual's emotional well-being. However, the specific implementation and analysis of these data will depend on the details of your machine learning pipeline and the goals of your depression detection model.



The screenshot shows a Windows File Explorer window with the address bar set to 'This PC > New Volume (D:) > Extracted'. The ribbon includes icons for Back, Forward, Copy, Paste, Delete, Sort, View, and a More options menu. The main area displays a list of folders named 300_P through 312_P, each with a date and time of modification and a type of 'File folder'.

Name	Date modified	Type	Size
300_P	01-09-2023 09:53	File folder	
301_P	01-09-2023 09:53	File folder	
302_P	01-09-2023 09:53	File folder	
303_P	01-09-2023 09:53	File folder	
304_P	01-09-2023 09:54	File folder	
305_P	01-09-2023 09:54	File folder	
306_P	01-09-2023 09:54	File folder	
307_P	01-09-2023 09:54	File folder	
308_P	01-09-2023 09:54	File folder	
309_P	01-09-2023 09:54	File folder	
310_P	01-09-2023 09:54	File folder	
311_P	01-09-2023 09:54	File folder	
312_P	01-09-2023 09:54	File folder	

Fig 3.4.1 File Structure with Patient wise folder



Name	Date modified	Type	Size
300_AUDIO	14-08-2023 12:43	WAV File	20,266 KB
300_CLNF_AUs	14-08-2023 12:43	Text Document	2,126 KB
300_CLNF_features	14-08-2023 12:43	Text Document	23,493 KB
300_CLNF_features3D	14-08-2023 12:43	Text Document	35,626 KB
300_CLNF_gaze	14-08-2023 12:43	Text Document	2,917 KB
300_CLNF_hog	14-08-2023 12:43	Text Document	3,39,603 KB
300_CLNF_pose	14-08-2023 12:43	Text Document	1,637 KB
300_COVAREP.csv	14-08-2023 12:43	Excel.CSV	35,954 KB
300_FORMANT.csv	14-08-2023 12:43	Excel.CSV	2,100 KB
300_TRANSCRIPT.csv	14-08-2023 12:43	Excel.CSV	9 KB

Fig 3.4.3 Individual files under each Patient

Chapter 4: Proposed Solution

Feature Extraction: In our system, we plan to first extract features and then apply some gating mechanism and hybrid fusion technique on the features extracted. For feature extraction, we have audio, visual, and textual modalities as our features that are integrated using time-stamps to learn the time-dependent interactions between them. The forced alignment will be done on a sentence level granularity. This is because we want the model to learn the context between words. This is the preprocessing part.

Alignment of Modalities: Now, we have aligned the textual, audio, and visual features at the sentence level. One important thing to note is that different modalities can have different impacts on the final result, and there is some noise involved too while representing the features of different modalities.

Gating Mechanism: On the extracted features, some gating mechanism will be applied to learn and control how much different modalities will be contributing to the final output. In our network, we'll use weight vectors with each modality to control and learn how much information will be transformed and carried to the next layers. For each time step, the feature vectors from each modality will be concatenated and then passed to the word-level LSTM, which comprises the gating mechanism. Before the concatenation of the feature vectors, the audio and visual vectors will also be passed through a gating mechanism to extract the important information.

Hybrid Fusion Technique: The other approach that we can follow is to use a hybrid fusion technique to reap the benefits of both early and late fusion. Hybrid fusion can be performed on one level or two levels. In hybrid fusion, feature fusion of all modalities is performed first to create a new modality, which is then treated as an additional individual modality. The scores/decisions of this new modality are then fused with the scores/decisions of the individual modalities in either one or two levels.

Chapter 5: Work Done and Results

5.1 CNN Model Architecture:

A Convolutional Neural Network (CNN) model was constructed with a total of 6 layers. The architecture differs based on the modality being processed.

For Text Modality:

- The first 4 layers consist of Conv2D layers.
- The Word2Vec model is applied to process textual data.
- Thresholds are set for the maximum number of words and sentences.
- ReLU activation function is used in intermediate layers.
- Sigmoid activation is employed in the final layer.

For Audio and Video Modality:

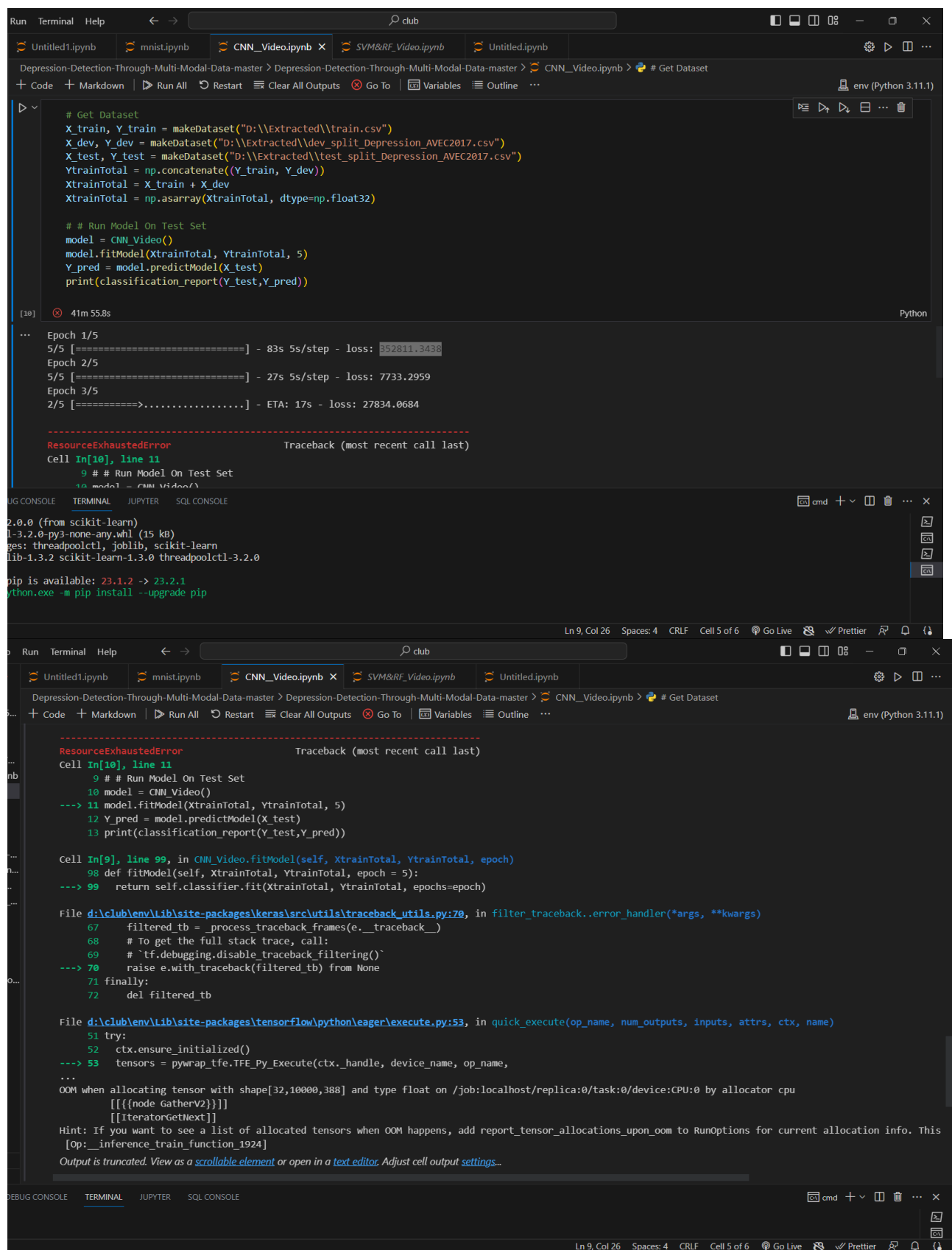
- The first 4 layers include Conv1D layers.
- Max Pooling layers are added for down-sampling.
- ReLU activation is applied in intermediate layers.
- Sigmoid activation is used in the last layer.

Modality-Specific Parameters:

- For the audio modality, features from the first 40,000 timestamps are considered.
- For the video modality, timestamps are utilized.
- The selection of these values takes into account the available computation capability and specific requirements of each modality.

This CNN model architecture is designed to process different modalities efficiently, extracting relevant features and making predictions, with the choice of parameters tailored to the characteristics of each modality.

Outputs:



The image displays two screenshots of a Jupyter Notebook interface, likely JupyterLab, showing the execution of a Python script and the resulting error.

Top Screenshot: The notebook is titled "CNN_Video.ipynb". The code cell shows the following steps:

- Get Dataset: Load training and testing data from CSV files.
- Run Model On Test Set: Train a CNN model and predict on the test set.

The output shows the training progress for 5 epochs:

```
Epoch 1/5
5/5 [=====] - 83s 5s/step - loss: 352811.3438
Epoch 2/5
5/5 [=====] - 27s 5s/step - loss: 7733.2959
Epoch 3/5
2/5 [=====>.....] - ETA: 17s - loss: 27834.0684
```

A **ResourceExhaustedError** is raised, indicating an Out Of Memory (OOM) error. The error message is:

```
ResourceExhaustedError: Traceback (most recent call last)
Cell In[10], line 11
      9 # Run Model On Test Set
     10 model = CNN_Video()
     11 model.fitModel(XtrainTotal, YtrainTotal, 5)
```

Bottom Screenshot: This screenshot shows the full traceback of the error, starting from the `fitModel` method in the `CNN_Video` class, through the `filter_traceback.error_handler` function, and finally to the `quick_execute` function in the `tensorflow/python/eager/execute.py` file. The error occurs when allocating a tensor with shape `[32, 10000, 388]` and type `float` on the device `GPU:0` by the `allocator cpu`.

The bottom screenshot also shows the `DEBUG CONSOLE` tab, which displays the error details and the `Output` tab, which shows the truncated output of the cell.

References

- [1] 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 the 3rd ACM international workshop on Audio/visual emotion challenge, ACM, 2013, pp. 21–30.
- [2] Z. Yu, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell, "Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs," in Semdial 2013 Dial-Dam: Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue, 2013, pp. 160–169.
- [3] Y. Gong and C. Poellabauer, "Topic modeling based multi-modal depression detection," in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, ACM, 2017, pp. 69–76.
- [4] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. 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, ACM, 2017, pp. 61–68.
- [5] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio-based depression classification," in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016, pp. 35–42.
- [6] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016, pp. 43–50.
- [7] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in Proc. Inter-speech, 2018, pp. 1716–1720.
- [8] [Link to a PDF document.](#)
- [9] University Of Southern California. "DAIC-WOZ Dataset." [Dcapswoz.ict.usc.edu](http://dcapswoz.ict.usc.edu), dcapswoz.ict.usc.edu.
- [10] Alghowinem, Sharifa Goecke, Roland Wagner, Michael Epps, Julien Hyett, Matthew Parker, Gordon Breakspear, Michael. (2016). Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors. IEEE Transactions on Affective Computing. PP. 1-1. [DOI: 10.1109/TAFFC.2016.2634527](https://doi.org/10.1109/TAFFC.2016.2634527).
- [11] Rohanian, Morteza Hough, Julian Purver, Matthew. (2019). Detecting Depression with Word-Level Multimodal Fusion. [DOI: 10.21437/Interspeech.2019-2283](https://doi.org/10.21437/Interspeech.2019-2283). PP. 1443-1447.