National College of Ireland

Multimodal Depression Detection using Audio and Visual Features

MSc Research Project
MSc Artificial Intelligence

Sachleen Singh Chani
Student ID: x22244778

School of Computing
National College of Ireland

Supervisor: Dr. Muslim Jameel Syed

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sachleen Singh Chani
Student ID:	x22244778
Programme:	MSc Artificial Intelligence
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr. Muslim Jameel Syed
Submission Due Date:	12/08/2018
Project Title:	Multimodal Depression Detection using Audio and Visual Features
Word Count:	5965
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sachleen Singh Chani
Date:	16th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

List of Figures

List of Tables

Multimodal Depression Detection using Audio and Visual Features

Sachleen Singh Chani
x22244778

Abstract

Depression is one of the most common mental health disorders, yet the diagnosis for this is either not readily available or often misdiagnosed. Recent studies using machine learning techniques for depression detection have shown promising results but more research needs to be done in using multiple modalities of data for depression detection. To address this, this research outlines a BiLSTM with Attention layer model with two separate pathways, one for audio modality and the other for video modality. Further, a comparison is done with single modality models to evaluate the proposed model. The testing was done on the DAIC-WOZ dataset and the porposed model was able to achieve an accuracy of 0.934 and recall of 0.944.

1 Introduction

Depression significantly affects the normal day-to-day activity of individuals (Friedrich; 2017). It is a common mental health condition and a research estimated that around 7% adult population in the US have depression [1]. Yet at present the most commonly used method for diagnosing depression is through clinical interviews and questionnaires. For which the validity of the diagnosis depends on the cooperation of the patient (Xue et al.; 2024). Utilization of machine learning in the process of detecting depression can provide a better and more accurate outcome. Further the implementation of these machine learning systems would make the diagnosis process readily available.

The use of machine learning techniques have gained a lot a attention in the recent years. Majority of these systems mainly utilize single modality of the data, either just textual, or through speech. And as compared to the unimodal systems, multimodal approach provided a more comprehensive understanding of the depression-related cues for these automated depression detection systems (Alghowinem et al.; 2018). Through this, the bidirectional long short-term memory model, which processes data in forward and backwards direction to find dependencies, it is particularly beneficial in case the data has a temporal aspect.

More and more complex models had been used for this task of depression detection, which are computationally taxing for training and often require considerable resource for implementation. To address this issue, this research aims at increasing the performance of the multimodal classification systems using BiLSTM along with attention mechanism. One main focus for the evaluation is to keep the number of false negative predictions low, since predicting a patient is not-depressed while they are, is vastly more dangerous.

[1] https://my.clevelandclinic.org/health/diseases/9290-depression

1.1 Research Question

How to improve the performance of algorithms for the binary classification of depression detection using Multimodal approach through a less computationally taxing method. The main focus for the improvement is on reducing the number of false negative predictions.

Section 2 provides a detailed review of the current literature and the research conduction in the field of depression detection using various machine learning algorithms. Section 3 discusses the dataset used, feature engineering, data preprocessing, a brief explanation unimodal models investigated and the proposed model for this research, and the evaluation metrics used. Followed by a detailed description of the machine learning and deep learning models outlined in section 4. The implementation of the proposed model is discussed in section 5, followed by the evaluation and comparison of the models in section 6. Finally the conclusion is presented in section 7 along with future research scope.

2 Related Work

There has been multiple research conducted on the topic of mental health disorders. A variety of traditional machine learning algorithms and deep learning have been applied for depression classification.

2.1 Uni-Modal Models

Miao et al. (2022) proposed a novel fusing higher-order spectral features (HOSA) with the traditional COVAREP audio features. The dataset used for this study is the DAIC-WOZ dataset, which comprises of the audio features. The authors extensively test the SVM, KNN and CNN model on this fusion of features, concluding that the fusion of the COVAREP features with HOSA improves the capability of these models to recognize the depression symptoms from the features. The CNN model outperforms other tested models with the accuracy of 85%. Though the research lacks in testing more deep learning algorithms proven to perform better with sequential data. This limitation of model's ability to capture depression-related cues was addressed by the use of graph neural networks in the study conducted by Sun et al. (2024). They focused on the audio features, and DAIC-WOZ, D-Vlog anf MODMA datasets were leveraged to train their model. The use of gated recurrent unit (GRU) was able to extract relevant time-series features from the audio. By the implementation of self-attention mechanism, they were able to achieve an accuracy of 61.64% with training of MODMA dataset and testing on DAIC-WOZ dataset. Although the multiple dataset was incorporated in their study, the data characteristics would differ significantly, posing difficulty in the predictions. They have suggested that more research is needed to address this issue.

Similar research was conducted through the MFCC audio features by Das and Naskar (2024). They proposed a novel CNN architecture for the classification on the audio spectogram features and the MFCCs. The proposed Spectro_CNN fuses the MFCCs and the acoustic patterns (spectrogram) on the participants voice data. Through the implementation of Leaky ReLU to mitigate the problem of dead neurons, their study was able to achieve an accuracy score of 0.896 and a recall of 0.914. Though, the limitation of single modality still exists, there also exists a high dependence of the model's performance on the quality of the audio data.

Working with the speech data, and leveraging various machine learning and deep learning techniques, the study by Kanoujia and Karuppanan (2024) used SVM and random forest along with CNN and RNN to capture the intricate speech patterns. Their main focus is to identify speech characteristics which highlight depression symptoms. By leveraging early stopping and hyperparameter tuning, the CNN model was able to achieve an accuracy of 68.83%, while SVM giving the highest accuracy of 0.772, outperforming the CNN.

Yalamanchili et al. (2020) developed a real-time depression classification system based on an android application through the use of textual features from the DAIC-WOZ dataset. The focus of their study was the extraction of prosodic, spectral and voice control features for the classification task. The SVM model trained on these three fused features was able to achieve an accuracy of 93%. Similarly, using the uni-modality text features, and the use of BiLSTM with self attention in the study conducted by Li et al. (2022). Their study compared the results between considering the automated interviewer 'Ellie' questions along with the participants responses and without 'Ellie' questions. They reported higher f1 score when the questions from 'Ellie' were considered along with the participants responses.

Through the use of only facial features, particularly the action unit (AUs) features from the DAIC-WOZ dataset, in their study (Akbar et al.; 2021) was able to get an accuracy of 97.83%. They leveraged the use of FFNN and was trained on the AUs features. The choice for FFNN is due the effectiveness and the reduced computation that these networks provide. They investigated the performance of this network for various training epochs, hidden nodes, and learning rates and concluded that the use of particle swarm optimization was able to improve the performance along with the use of backpropogation training technique. Through their testing, they were able to achieve an accuracy of 97.83% though the use of bayesian regularization.

2.2 Multi-modal Models

Multi-modality for the task of depression detection introduces more features the models can train on, thus improving the performance and effectiveness. Through the fusion of selected features like paralinguistic, head pose and eye gaze, the study conducted by Alghowinem et al. (2018) leveraged the DAIC-WOZ dataset and used SVM classifier. They concluded that the use of complementary multiple modalities can enhance the prediction performance. With their reaserch concluding with an accuracy of 88% while accuracy for individual modalities were 83% for speech, 73% for eye and 63% for head. Their research also delved into the misclassifications by their model. One reason they concluded was the misclassification of males more than that in females.

Another issue with depression detection is the absence of diverse dataset. The patients might be from different regions, or speak different languages. The Collection and evaluation on the EATD-corpus introduced a public dataset for depression detection (Shen et al.; 2022).

In their study (Shen et al.; 2022), they tested on the EATD-Corpus and DAIC-WOZ datasets. They present a multimodal classificaion model leveraging both the speech characteristics and linguistic content from the interviews. They complied the Emotional Audio-Textual Depression (EATD) corpus, this filled an important gap in the absence of foreign language datasets for depression detection. Being the first of its kind dataset in China gives more avenues for further research. The proposed model in this study utilized

the fusion of audio and text data, leveraging multi modalities which demonstrated better generalization with the f1-score of 0.71 and recall value of 0.84.

In the study by Yang, Sahli, Xia, Pei, Oveneke and Jiang (2017) they deployed a hybrid model for the depression classification task, implementing on the DAIC-WOZ dataset. The hybrid framework integrated CNN, deep neural network, paragraph vectors and support vector machines. Their proposed approach leverages both unimodal and multimodal strategies for the prediction of the PHQ8 score, a regression classification task. Through the evaluation the research was able to reduce the root mean square error (RMSE) and mean absolute error (MAE), with root mean square value of 5.40 and mean absolute value of 4.36. They utlizied a late-fusion approach, where they initial predictions from the two modalities were fused through a deep neural network. Similar hybrid approach was done in the research conducted by Saidi et al. (2020) in which they combined CNN with SVM. Through this model the features were automatically extracted via the CNN and the SVM was the classifier. The training and evaluation was done on the same DAIC-WOZ dataset, achieving an accuracy of 68%.

To address the issue of limited data, Yang, Jiang, Xia, Pei, Oveneke and Sahli (2017) incorporated the use of textual modality along with the audio and video modalities in another study by them utilizing multimodality regression. Their proposed fusion-model used deep CNN to learn high-level global features while the deep neural network was the predictor for the PHQ8 score. The PHQ8 value for the three modalities were predicted and then integrated through the deep neural network for the final prediction. A novel appraoch was taken for the video features by calculating the Histogram of Displacement Range (HDR) from the 3D and 2D facial features. Their proposed model was able to obtain a root mean square value of 4.65 and mean absolute error of 3.98.

BiLSTM has proven to be better suited for processing of sequential data. Iyortsuun et al. (2024) proposed the use of BiLSTM for the audio and video features for their multimodal approach. The testing and evaluation was done on the DAIC-WOZ dataset. Their model was able to achieve a f1-score of 0.82 wit precision and recall of 0.79 and 0.86 respectively. To tackle their limitations, they suggested a better solution for the class imbalance that is present in most of the depression dataset.

Building upon the use of target approach towards each modality, (Xue et al.; 2024), in their study leverage the use of text features through text sentence embedding and audio features through multi-level audio features interaction module (MAFIM). The use of pretrain BERT model to extract sentence level embeddings provided a more effective methods of extracting contextual information from the interview text data. For fusing the predictions from the text and audio modality they designed a channel attention-based multimodal fusion module (CAMFM). The evaluation of their proposed model on the DAIC-WOZ dataset produced 0.92 f1 score, 0.92 recall and 0.92 precision.

Attention layers can significantly improve the performance of the classifiers by focusing on the most relevant features during training. This implementation of attention based CNN model was implored in the study by Tiwary et al. (2023). In their research they implemented two A-CNN models for the audio and video modality and concatenated them though the use of dense layers. This proposed model was trained on the DAIC-WOZ dataset and was able to obtain an accuracy value of 0.771 with precision and recall value of 0.667 and 0.667 respectively. They reported a higher f1 score of 0.826 for non-depressed participants. One reason for this higher level of f1 score could be due to the non-depressed cases being a majority class in the DAIC-WOZ dataset.

By incorporating all three modalities namely audio, video and text data present in the

DAIC-WOZ dataset, the study by (Patapati; 2024) leveraged MFCC for audio modality, facial Action Unit features from the video modality and the novel appraoch of the utilization of GPT-4 model for the text modality. Through their research their proposed model was able to achieve an accuracy of 0.9101 and a recall of 0.9286 with f1 score of 0.8595. Though their model was able to achieve impressive results, there are limitations to their approach. Primary limitation is due to the use of large language model like GPT-4 which makes this application non-viable for real-time deployment. Further the use of limited data for their study also limits their model's generalizability.

The previous research solely focused on the data to train their proposed models, to move away from purely data-driven methods, Yang et al. (2024) approached the depression classification through a novel method called Multimodal Purification Fusion, they proposed the use of audio and video data for a better depression detection by incorporating the use of prior constraint gating (PCG), which uses the knowledge from clinical experts to help the model to focus on relevant psychological features, and the use of enhanced feature extraction along with fusion through transformers. Through their research they were able to achieve a recall of 0.93 and accuracy of 0.87 with an f1 score of 0.88. Their approach despite achieving high performance, the computational complexity is high considering the utilization of PCG and multimodal integration through transformers.

3 Methodology

In this section, a detailed description of the approach used to build and evaluate the multimodal depression detection model. For this research, along with the multimodal model, unimodal models are also investigated through several deep learning and machine learning model to evaluate their performance and contrast them with the multimodal model.

The main focus of this research is to investigate auditory and visual features for the purpose of depression detection.

3.1 Dataset

The data used for this research is from Distress Analysis Interview Corpus - Wizard-of-Oz (DAIC-WOZ) dataset (Gratch et al.; 2014). This dataset is a collection of clinical interviews conducted by the University of South California. This dataset was specifically curated to aid in the diagnosis of mental disorders and was introduced in the AVEC2017 Challenge (Ringeval et al.; 2017). It contains data from 189 participants and was collected through clinical interviews, that were conducted in Wizard-of-Oz style using a virtual interviewer called Ellie, which was controlled by a human interviewer from another room. The reason for the choice of this data is due to it being a standard dataset upon which various studies have conducted. This help further to contrast the performance of the proposed model.

This data contains audio, video and Patient Heath Questionnaire (PHQ8) responses in text format for each participant. For ethical reasons and consent constraints, the video file is not provided, rather extracted features are provided. Features extracted from the interview audio is also included in the dataset.

For this research, the audio and video features were leveraged for the training of the proposed model from this dataset. Several of the audio and video datapoints were not suitable for the training due to interruptions during the interview or technical issue. These

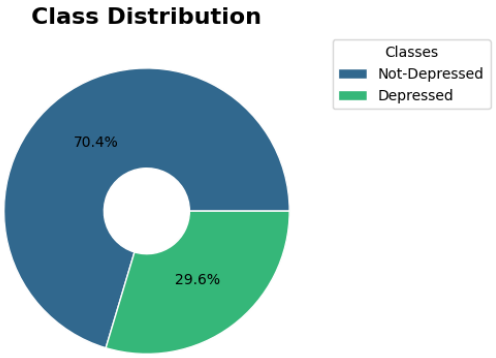

Figure 1: Depressed and not-depressed class distribution

Table 1: COVAREP Feature Description

Feature	Description
F0	Pitch of the voice (Hz)
VUV	Voiced or Unvoiced audio segment. Voiced segments have vibrating vocal cords
NAQ	The ratio of the harmonic components to the noise components
QOQ	Duration the vocal cords are open during a cycle of vibration
H1H2	Ratio of the first harmonic to the second harmonic
PSP	Peak-to-slope ratio of the voice signal
MDQ	Measure of jitter, variability in the pitch
peakSlope	Indicates the slope of the peak in the voice signal
Rd	Measure of the signal's pitch from the expected pitch
Rd_conf	The confidence of the Rd measure
MCEP_0-24	Mel-Cepstral Coefficients
HMPDM_0-24	Harmonic Parameters of the Deviation from the Mean
HMPDD_0-12	Measure the duration of Harmonic Parameters of the Deviation from the Mean

values were removed to maintain the integrity of the data for training and evaluation. The gender of the participants was also not considered for this research so as not to introduce gender bias.

There exists a class imbalance in the data, with majority data for non-depressed and depressed being the minority class. Fig. 1 shows the distribution of depressed and not-depressed participants. This was addressed using the Synthetic Minority Oversampling Technique (SMOTE).

3.1.1 Audio Feature

The audio features were extracted using the COVAREP toolbox (Degottex et al.; 2014). Table 1 gives a detailed overview of the features that are leveraged for this research.

Further, first five voice formants were also included in the research. Voice formants help to distinguish between the various speech patterns through specific peaks in the frequency spectrum of the voiced signal (Broad; 1972). Table 2 gives a list of the formants that were extracted.

A detailed overview of the COVAREP toolbox are available on the COVAREP github repository [2].

[2]https://github.com/covarep/covarep

Table 2: Voice Formants Description

Feature	Description
F1	Lowest resonant frequency of the vocal tract
F2	Second lowest resonant frequency
F3	Third lowest resonant frequency
F4	Fourth lowest resonant frequency
F5	Fifth lowest resonant frequency. Also contributes towards the timbre of the voice

Table 3: Visual Action Units

Face Section	AUs	Description
Upper Face	AU01	Inner eyebrow raiser
	AU02	Outer eyebrow raiser
	AU04	Eyebrow lowerer
	AU05	Upper eyelid raiser
	AU06	cheek muscle raiser
	AU45	Eyelid blinking
Lower Face	AU10	upper lip raiser
	AU12	lip corner puller
	AU15	lip corner depressor
	AU09	nose wrinkler
	AU14	dimpler
	AU17	chin raiser
	AU20	lip stretcher
	AU25	lips part
	AU26	jaw drop
	AU11	nasolabial deepener
	AU15	lip corner depressor
	AU23	lip tightener
	AU28	lip suck

3.1.2 Video Features

The video features were extracted using the CLNF framework (Baltrušaitis et al.; 2016). It provides a comprehensive set of facial features which gives a detailed understanding of the facial expressions, the gaze, and head position. This second modality, audio features would give meaningful insights into non-verbal cues. Due to ethical constraints the raw video file or the participants is not included in the dataset.

The features considered for this research are the 2D and 3D facial landmark. The Action units which capture the activation of various facial muscles would give insight into micro expressions and changes in expression, Table 3 shows a list of Visual Action units in the dataset.

Gaze features were also extracted, which give where the individual is looking, this includes the vector values in global and head-relative coordinates. Further the head's spacial orientation feature was also considered.

There are 20 AUs features, 68 2D and 3D facial landmark features, 12 gaze features, and 6 pose features.

These features give valuable insights into the subjects mental state, as a depressed individual has decreased eye contact, decreased eyebrow movement and an increase in the blinking rate (Alghowinem et al.; 2018).

3.2 Data Preprocessing

The video features extracted namely, gaze, pose, 2D and 3D face landmarks, and AUs, these are combined into datasets for each participant. The data points with low confidence and null values are removed. All the data was compiled into a video dataset. Fig. 2(a) shows the flow of the data with csv files created for each participant. Similarly, the audio features were preprocessed, to compile the audio dataset, fig. 2(b) depicts the audio data

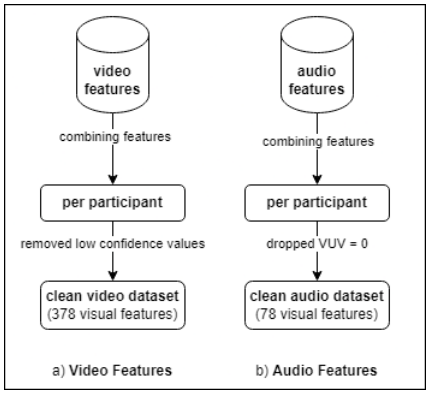

Figure 2: Data Preprocessing

preprocessing.

A notable difference in this research is leveraging an independent train-test split, rather than the utilization of the predefined training, testing and development split from the DAIC-WOZ dataset. This research has opted to explore the proposed model's performance and ability to generalize beyond the specific split provided, which would provide crucial for real-world data. This would highlight the model's strengths and weaknesses in varied contexts.

During the split, the distribution of the classes is maintained. Further, focus is also given to the sequential nature of the video and audio data.

3.3 Models

Various machine learning and deep learning models are leveraged for the binary classification task for depression detection. The unimodal model performance is also investigated on video features and audio features separately. This highlights the individual modality's contribution towards the classification task along with giving a baseline for the comparison with the multimodal model.

Random forest (RF), Light Gradient Boosting Machine (LightGBM) and Feed Forward Neural Network (FFNN) is utilized as baseline models for this research. They provided a diverse comparison as they cover both machine learning techniques and deep learning techniques, which allows for a more comprehensive evaluation against the proposed BiLSTM with attention model. The tree-based models like LightGBM and Random Forest can easily process tabular data with high dimensionality datasets which also makes them excellent at identifying nuanced patterns in the tabular data. Further, the use of deep learning architecture like FFNN gives insight into the performance of video modality, with the neural network's capability to capture non-linear complex relationship between the features and the target. The choice of random forest and LightGBM also aligns with popular methodology for depression detection since these models have shown promising results.

These models provide useful insight into how well non-sequential models like tree-based models and FFNN perform on the video and audio modalities. This collectively establishes a comprehensive baseline for a more complex deep learning approach like the BiLSTM architecture.

3.3.1 Unimodal Models

Unimodal models are trained and evaluated on a single modality data, for this research the audio and video modality are considered.

Random Forest (RF):

This tree based model, which creates and ensemble of decision trees. This machine learning model is able to capture complex relations between the features and the target. By leveraging multiple decision trees, the risk of overfitting during training is reduced. Further the ability for the random forest model to parallelize the training, it can handle vast amounts of data. This makes it an ideal base line model for this research. This model is trained on the video modality data due to the high number of features (378 features).

Light Gradient Boosting Machine (LGBM):

This is another tree based model, which provides further methods in preventing overfitting which is, feature randomization. At each tree split the model selects a random subset of features, this reduces the correlation between the constructed trees. The model provides much high efficiency during training, thus making it a optimal choice for this research.

Feed Forward Neural Network (FFNN):

This model provides a straightforward method for training, and requiring less computation. This helps to target the required hyperparameter tuning for the video modality of the multimodal model.

3.3.2 Multimodal Model

Bidirectional Long Short-Term Memory (BiLSTM) with Attention:

Due to the bidirectional aspect of this model, it makes this model highly efficient with sequential data. The model's ability to leverage forward and backward direction of processing the sequential data makes the model effective for this research. This ability to capture details is further increased by the use of an attention layer, which focuses on the most relevant parts of the temporal data.

3.4 Evaluation Metrics

For the evaluation of the models experimented for the research, various metrics have been used namely, accuracy, precision, specificity, f1-score, and ROC-AUC. Along with these, more focus is given to metrics like recall(sensitivity), Negative predictive value (NPV), Matthew's correlation Coefficient (MCC) and False negative Rate (FNR).

Since the task for binary classification is for depression detection, although the false positive predictions are not desirable, higher importance is given to false negative predictions, since predicting not-depressed for a depressed case is more harmful, thus making it critical to minimize the false negative predictions.

False Negative Rate (FNR):

This metric directly measure the rate at which the model is failing to predict the true cases of depression, which it is classifying falsely as negative predictions. A lower value of FNR indicate the model is able to capture the depressed cases correctly.

Negative Predictive Value (NPV):

NPV gives the ratio of correctly predicting the not-depressed cases. Higher value of Negative predictive value means that if the model predicts a person is not-depressed,

there is a higher chance that the model is correct. Thus minimizing the false negative predictions.

Sensitivity (Recall):

This metric gives the evaluation of the model's ability to correctly identify as many true cases of depression as possible, highlighting the chance of a false negative prediction. A higher value indicates the model is able to correctly identify true cases of depression.

Matthew's Correlation Coefficient (MCC):

Although Matthew's correlation coefficient doesn't directly evaluates the false negative cases, rather it gives a balanced evaluation of the model by considering all true positive, true negative, false positive and false negative values. A higher MCC value indicates the model's overall performance is high.

Furthermore, Receiver Operating Characteristic (ROC) curve is also plotted for the models to evaluate their performance at all classification thresholds.

The following section will delve further into details of the design of the models. And further sections will evaluate the implemented models and the proposed multimodal model based on the metrics discussed.

4 Design Specification

This section will detail the architecture of the unimodal and multimodal models that are implemented. The proposed multimodal model integrates the audio and video modalities to achieve effective classification. Following subsections describe the models and frameworks used in this reserach.

4.1 Data Processing

The dataset used for the training of the models is the DAIC-WOZ dataset. After the preprocessing steps outlined in the previous section, the data is split into target and features for both the audio and video modalities.

4.2 Random Forest

Random forest model combined multiple decision trees through ensemble which improves the classification performance. For this research this model is used on the video modality data to create a baseline for evaluation. The target the features were split before the model is trained on the video features. The number of trees utilized is 100 for the classification.

4.3 Feed Forward Neural Network (FFNN)

The network is constructed with an input layer with 378 nodes to capture the input feature dimension of the video extracted features. Subsequently, 5 hidden layers are implemented with Relu activation and finally an output layer with a single node with sigmoid activation for the binary classification. Fig. 3 shows the diagram for the neural network.

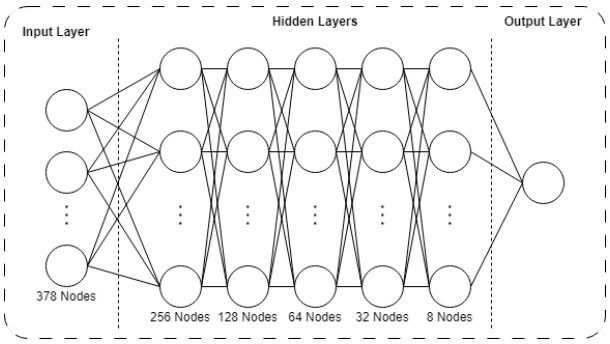

Figure 3: FFNN Diagram

4.4 Light Gradient Boosting Machine

This tree based model was given the same train-test split as all the other unimodal models so the models can be better evaluated on their performance with a uniform data-split amongst them. For simplicity, the hyperparameters are set to default. The random state is set to 42 for repeatability of the experiments.

4.5 BiLSTM + Attention (unimodal)

This model was implemented on the audio feature dataset. The BiLSTM processes the sequential data forwards and backwards, this comprises of 64 units. Further an attention layer is used to capture the significant trends of the audio features. A POoling layer is applied to reduce the dimensionality by applying global max pooling, and finally a fully connected dense layers for the binary classification.

The proposed multimodal BiLSTM with attention model is discussed in detail in the following Implementation section.

5 Implementation

5.1 BiLSTM + Attention (Multimodal)

The focus of this research is the proposed Bidirectional Long Short-Term Memory model with attention. This model concatenates the two separate BiLSTM networks, each for the audio and video modalities. The merging of the models is done after the implementation of separate attention layer to focus the BiLSTM model classifications on the two modalities.

Fig. 4 gives a high-level view of the flow of this research.

The processed audio features extracted from the DAIC-WOZ dataset are reshaped, similary the video features are reshapes, keeping close attention to obtain a consistent shape on both the inputs to maintain the sequential aspect for both the modalities.

Following this, the audio and video features are split from the target class. The audio features are fed into the audio branch of the BiLSTM model, while the video features are fed into the video branch. The outputs from these each are separately concatenated with an attention layer and global max pooling is applied to reshape the outputs. These audio and video pathways are then concatenated together and two dense layers are applied for

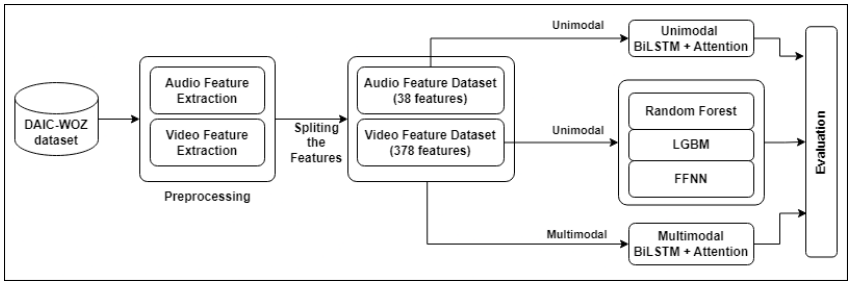

Figure 4: High-level Design Implementation

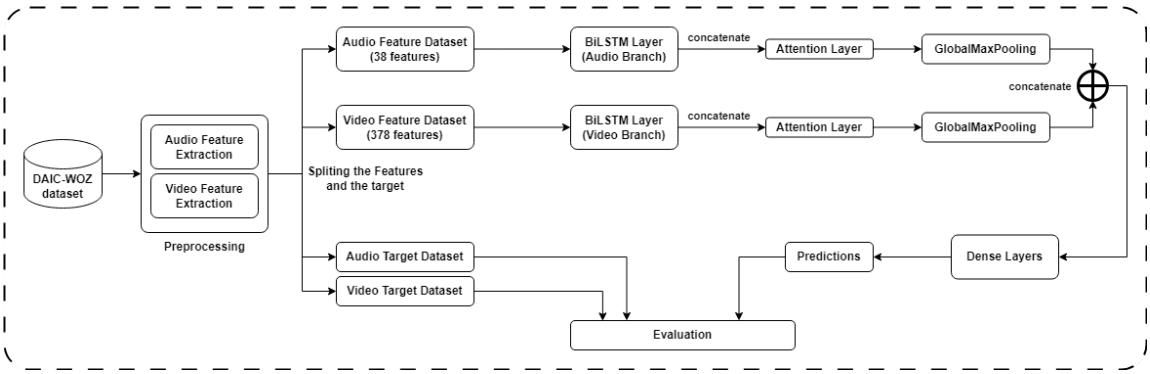

Figure 5: Multimodal BiLSTM with Attention

the binary classification. The output layer has a single node with sigmoid activation. This design can be seen in Fig. 5.

This model is evaluated on the previously defined metrics, while giving more focus on the false negative classifications.

6 Evaluation

This section presents and scrutinizes the evaluation results. For this, the evaluation of the experiments conducted for this research is done through various metrics. The main focus is on the following metrics namely, recall, negative predictive value, false negative rate and Matthew's correlation coefficient. Along with these other metrics leveraged are accuracy, precision, specificity, f1-score, ROC curve and ROC-AUC.

The formulas for these metrics is as follows:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{P+N} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP}+\text{TN}} \tag{3}$$

$$\text{F1 Score} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \tag{4}$$

$$\text{False Negative Rate (FNR)} = \frac{\text{FN}}{\text{FN}+\text{TP}} \tag{5}$$

$$\text{Negative Predictive Value (NPV)} = \frac{\text{TN}}{\text{TN+FN}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{7}$$

$$\text{Matthews Correlation Coefficient (MCC)} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \tag{8}$$

6.1 Experiment 1

For this, a single modality is considered, i.e. video modality. The features extracted from the video modality are used to train Random Forest, LightGBM. Table 4 shows the gathered results from this experimentation. The main goal for this experiment was two fold, first was to construct a baseline using machine learning algorithms on a single modality for the evaluation of the proposed multimodal, other was to gain insight into the contribution of the video modality towards the final multimodal classification.

Table 4: Evaluation Metrics for Video Modality using Machine learning models

Modal	Accuracy	Precision	Recall	F1 Score	Specificity	NPV	ROC-AUC	FNR	MCC
Random Forest	0.775	0.888	0.312	0.462	0.982	0.761	0.647	0.688	0.437
LightGBM	0.719	0.644	0.214	0.322	0.947	0.728	0.581	0.786	0.245

From the results, The False negative rate for the LGBM is 0.786, which is higher than the 0.688 for random forest. This indicates that LGBM model is able to correctly capture the depressed cases. Although for both the models the overall performance is very low. The recall for both these tree based models is very less at 0.31 for random forest and 0.21 for LGBM, again suggesting that the model is not able to predict correctly for the non-depressed cases.

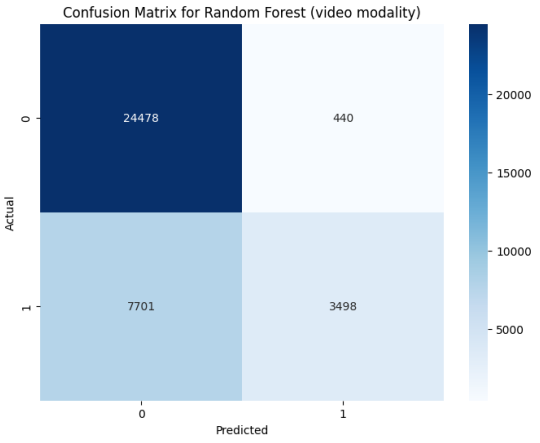

Figure 6: Random Forest (video modality) Confusion Matrix

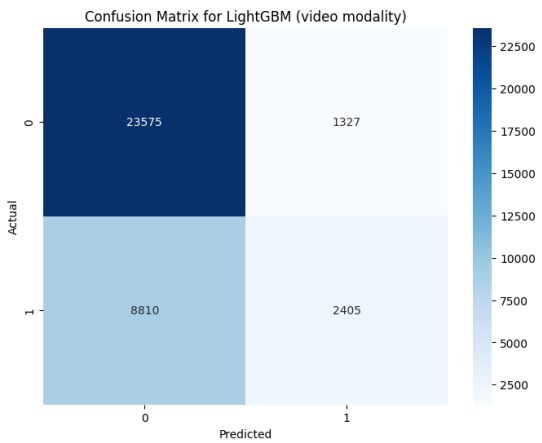

Figure 7: LGBM (video modality) Confusion Matrix

This can also be verified from confusion matrix for random forest in Fig. 6 and the confusion matrix for LGBM in Fig. 7. Both have very high false negative values at 7701 and 8810 respectively. Another point to highlight is the ROC-AUC values for these models, which is close to 0.5 which is for a random classifier. So the models predictions are slightly better than a random classifier.

6.2 Experiment 2

From the previous experimentation, it is clear that basic machine learning models are not able to capture the nuances from the video features. To further the research, more complex deep learning algorithms are applied, while keeping the same feature-target split for consistency. This would also verify if there was issue in data preprocessing or the models applied.

For this experimentation step, Feed forward neural network is trained on the video modality and Bidirectional LSTM with attention layer is trained on the audio modality separately.

Table 5: Evaluation Metrics Unimodal Deep learning models

Modal	Accuracy	Precision	Recall	F1 Score	Specificity	NPV	ROC-AUC	FNR	MCC
FFNN (video)	0.780	0.658	0.607	0.631	0.858	0.829	0.732	0.393	0.476
BiLSTM + Attention (audio)	0.877	0.841	0.720	0.776	0.943	0.889	0.832	0.280	0.696

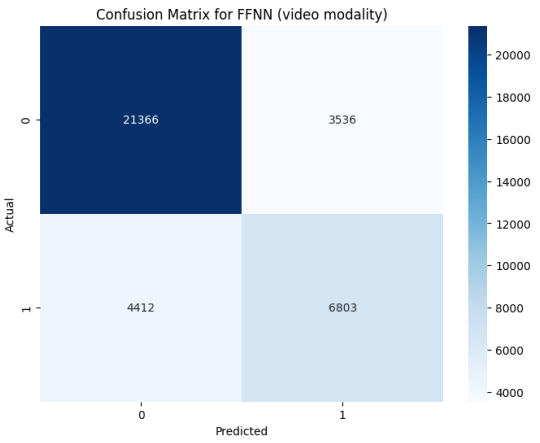

Figure 8: FFNN (video modality) Confusion Matrix

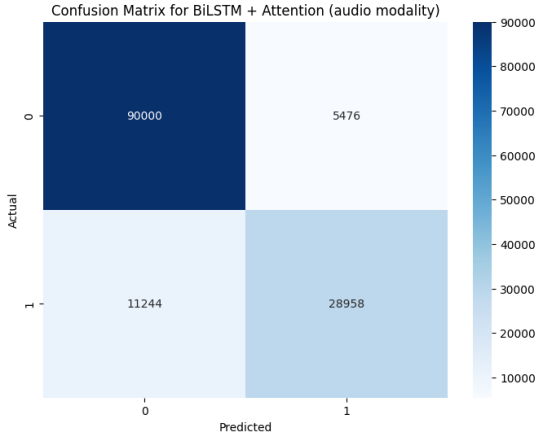

Figure 9: BiLSTM + Attention (audio modality) Confusion Matrix

The performance of these deep learning models are a slightly better than the machine learning algorithms. The metrics are shown in Fig. 5. Although more thorough investigation could be done with applying all four of these models on both the video and audio separately and then comparing.

Comparing the performance of FFNN and the BiLSTM + attention, we can see that the ROC-AUC value for these models are 0.73 and 0.83 respectively, with the BiLSTM having a higher 0.83 value. The overall results for the BiLSTM classifier is better when compared to the previous experiment and the FFNN model. The false negative rates for BiLSTM + attention is 0.28, which is lower than FFNN which is 0.39. This suggests that the bidirectional aspect of the LSTM along with the attention layer is able to capture features correlation with the target.

Looking at the precision-recall curve for FFNN in Fig. 10 and BiLSTM in Fig. 11 we can also confirm the performance of the BiLSTM is better, with a higher precision to recall balance than in the FFNN model. This experiment suggest that the performance of the BiLSTM model for the multimodal classification would provide better and efficient results.

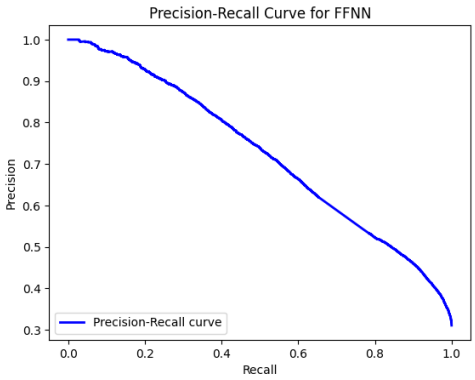

Figure 10: FFNN (video modality) Precision-Recall curve

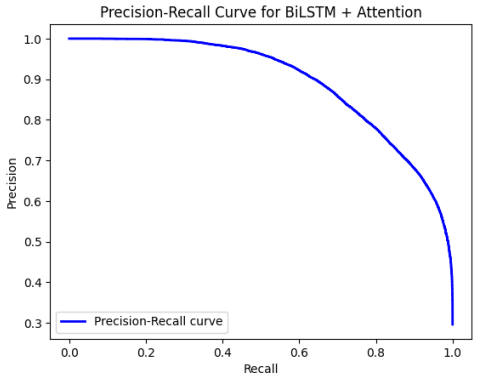

Figure 11: BiLSTM with Attention (audio modality) Precision-Recall curve

The following in the proposed model from this research for the task of binary classification of depression using multimodal approach.

6.3 Proposed Model

From the the proposed model for the multimodal classification is Bidirectional Long Short-Term Memory with Attention. This model was trained on the same previous feature and target splits to maintain consistency for comparison.

Table 6: Evaluation Metrics Multimodal BiLSTM with Attention

Modal	Accuracy	Precision	Recall	F1 Score	Specificity	NPV	ROC-AUC	FNR	MCC
BiLSTM with Attention	0.9348	0.8594	0.9448	0.9000	0.9303	0.9739	0.9375	0.0552	0.8539

The results for the proposed model are compiled in the Table 6. From the false negative rate we can see the value is much lower than the unimodal models, at 0.06. The recall with the value of 0.94, further the negative predictive value of 0.97, both of these values together suggest that the proposed model is able to correctly classify true positive cases, thus reducing the false negative predictions.

With the Matthew's correlation coefficient of 0.85, suggest that the overall performance of the proposed model is high.

From the ROC curve in Fig. 12, we can see that the proposed model is able to keep the false positive rate low and the true positive rate high, suggesting that the model is able to better capture the trends in the actually positive cases (depressed) and the actually negative cases (not-depressed). Furthermore, we see very low false negative predictions from the confusion matrix in Fig. 14.

6.4 Discussion

From the experimentation done in this research, the unimodal BiLSTM with Attention model was able to outperform random forest, lightGBM and FFNN. Thus this was chosen for the multimodel evaluation.

Since these models (random forest, LightGBM, FFNN and BiLSTM with Attention) are not applied to both the individual audio and video modality, a more accuracy com-

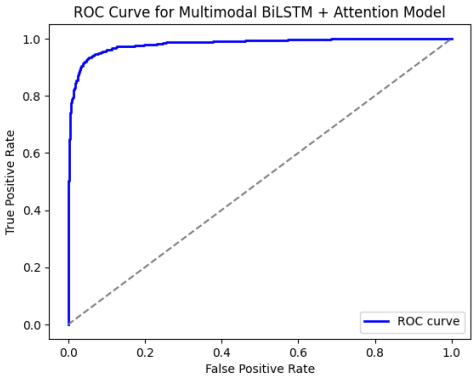

Figure 12: Multimodal BiLSTM with Attention ROC curve

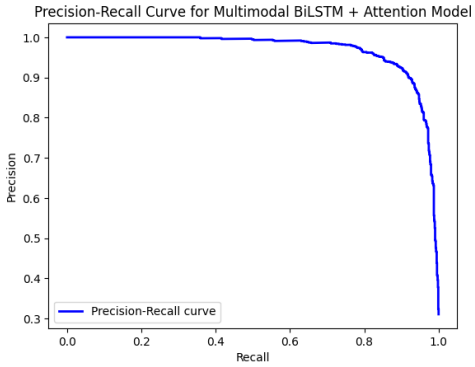

Figure 13: Multimodal BiLSTM with Attention
Precision-Recall curve

parison cannot be made. But the main focus when applying these models to the selected modality was to gain insight as well for, first, the performance of the models, second, the contribution of those modalities for the final multimodal classification. By conduction the above experiments that goal has been achieved.

Unimodal BiLSTM with Attention model that was applied to the audio modality was chosen and further applied for the multimodal classification, the reason being that the backwards and forwards processing of the bidirectional LSTM along with the ability of the LSTM algorithm to capture nuances in the sequential data. Further the implementation of the attention layer was able to focus the model training on the most relevant features.

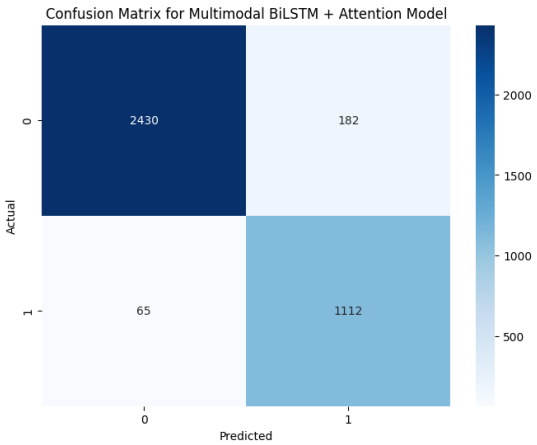

Figure 14: Multimodal BiLSTM with Attention Confusion Matrix

The proposed models performance can be compared with relevant research done through the use of same audio and video modalities. This result is compiled in the Table 7.

7 Conclusion and Future Work

This research explored the multimodal approach for the classification of depression. Through the experimentation, Bidirectional Long Short-Term Memory model with Attention mechanism was effectively able to capture trends in the audio and video modality.

Table 7: Evaluation of the proposed model with current literature

Method	Modality	Accuracy	Precision	Recall	F1 Score
Das and Naskar (2024)	Audio	0.896	**0.906**	0.914	**0.911**
Kanoujia and Karuppanan (2024)	Audio	0.77	0.75	0.77	0.71
Iyortsuun et al. (2024)	Audio + Video	-	0.79	0.86	0.82
Tiwary et al. (2023)	Audio + Video	0.771	0.667	0.667	-
Yang et al. (2024)	Audio + Text	0.87	0.83	0.93	0.88
Patapati (2024)	Audio + Video + Text	0.9101	0.80	0.9286	0.8595
BiLSTM + Attention (Proposed Model)	Audio + Video	**0.9348**	0.8594	**0.9448**	0.9000

This proposed model was evaluated and compared to the unimodal approach to assess the ability of the model to generalize beyond the pre-defined dataset split.

To achieve this, audio and video features extracted from the DAIC-WOZ dataset. Several machine learning and deep learning models, including Random Forest, LightGBM, and Feed Forward Neural Network were trained and evaluated on the single modality features. The proposed BiLSTM with attention model was trained and evaluated on the audio and visual modality features, and the performance was compared to the other models using various metrics.

The results from the research indicate that the BiLSTM with attention model was able to outperform the other models, particularly in key metrics like False Negative Rate, Recall and Matthew's Correlation Coefficient, with the false negative rate being 0.056, recall value of 0.94 and matthew's correlation coefficient value 0.85. The high recall and low false negative rate ensures that the model is able to correctly identify trends for the positive cases, thus recusing the false negative predictions in the context of depression detection, where incorrectly predicted true cases can have serious implications.

However, there are limitations of this research. This proposed model was only evaluated on a single dataset. While it performed well, a better generalization testing should be done for a wider demographic. Further the model has a very high dependence on the quality of the input data, any noise could affect the models performance.

For future research, additional modalities like textual can be incorporated to further enhance the performance. Another avenue would be to train the model on multiple datasets to create a more robust model. Furthermore, pre-trained models can used to extract features for each of the audio and video modality before concatenating would also be studied, this would help in real-world application research by reducing the computation required.

References

Akbar, H., Dewi, S., Rozali, Y. A., Lunanta, L. P., Anwar, N. and Anwar, D. (2021). Exploiting facial action unit in video for recognizing depression using metaheuristic and neural networks, *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, Vol. 1, pp. 438–443.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Hyett, M., Parker, G. and Breakspear, M. (2018). Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors, *IEEE Transactions on Affective Computing* **9**(4): 478–490.

Baltrušaitis, T., Robinson, P. and Morency, L.-P. (2016). Openface: An open source facial

behavior analysis toolkit, *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10.

Broad, D. J. (1972). Formants in automatic speech recognition, *International Journal of Man-Machine Studies* **4**(4): 411–424.
URL: *https://www.sciencedirect.com/science/article/pii/S0020737372800373*

Das, A. K. and Naskar, R. (2024). A deep learning model for depression detection based on mfcc and cnn generated spectrogram features, *Biomedical Signal Processing and Control* **90**: 105898.
URL: *https://www.sciencedirect.com/science/article/pii/S1746809423013319*

Degottex, G., Kane, J., Drugman, T., Raitio, T. and Scherer, S. (2014). Covarep - a collaborative voice analysis repository for speech technologies, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE.

Friedrich, M. (2017). Depression is the leading cause of disability around the world, *JAMA* **317**: 1517.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S. and Morency, L.-P. (2014). The distress analysis interview corpus of human and computer interviews, *in* N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 3123–3128.
URL: *http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf*

Iyortsuun, N. K., Kim, S.-H., Yang, H.-J., Kim, S.-W. and Jhon, M. (2024). Additive cross-modal attention network (acma) for depression detection based on audio and textual features, *IEEE Access* **12**: 20479–20489.

Kanoujia, S. and Karuppanan, P. (2024). Depression detection in speech using ml and dl algorithm, *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, Vol. 2, pp. 1–5.

Li, M., Xu, H., Liu, W. and Liu, J. (2022). Bidirectional lstm and attention for depression detection on clinical interview transcripts, *2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN)*, pp. 638–643.

Miao, X., Li, Y., Wen, M., Liu, Y., Julian, I. N. and Guo, H. (2022). Fusing features of speech for depression classification based on higher-order spectral analysis, *Speech Communication* **143**: 46–56.
URL: *https://www.sciencedirect.com/science/article/pii/S0167639322001029*

Patapati, S. V. (2024). Integrating large language models into a tri-modal architecture for automated depression classification.
URL: *https://arxiv.org/abs/2407.19340*

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M. and Pantic, M. (2017). Avec 2017: Real-life depression, and affect recognition workshop and challenge, *Proceedings of the 7th Annual Workshop on*

Audio/Visual Emotion Challenge, AVEC '17, Association for Computing Machinery, New York, NY, USA, p. 3–9.
URL: *https://doi.org/10.1145/3133944.3133953*

Saidi, A., Othman, S. B. and Saoud, S. B. (2020). Hybrid cnn-svm classifier for efficient depression detection system, *2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, pp. 229–234.

Shen, Y., Yang, H. and Lin, L. (2022). Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model.
URL: *https://arxiv.org/abs/2202.08210*

Sun, C., Jiang, M., Gao, L., Xin, Y. and Dong, Y. (2024). A novel study for depression detecting using audio signals based on graph neural network, *Biomedical Signal Processing and Control* **88**: 105675.
URL: *https://www.sciencedirect.com/science/article/pii/S1746809423011084*

Tiwary, G., Chauhan, S. and Goyal, K. K. (2023). Multimodal depression detection using audio visual cues, *2023 International Conference on Computer Science and Emerging Technologies (CSET)*, pp. 1–5.

Xue, J., Qin, R., Zhou, X., Liu, H., Zhang, M. and Zhang, Z. (2024). Fusing multi-level features from audio and contextual sentence embedding from text for interview-based depression detection, *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6790–6794.

Yalamanchili, B., Kota, N. S., Abbaraju, M. S., Nadella, V. S. S. and Alluri, S. V. (2020). Real-time acoustic based depression detection using machine learning techniques, *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pp. 1–6.

Yang, B., Cao, M., Zhu, X., Wang, S., Yang, C., Ni, R. and Liu, X. (2024). Mmpf: Multimodal purification fusion for automatic depression detection, *IEEE Transactions on Computational Social Systems* pp. 1–14.

Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M. C. and Sahli, H. (2017). Multimodal measurement of depression using deep learning models, *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, Association for Computing Machinery, New York, NY, USA, p. 53–59.
URL: *https://doi.org/10.1145/3133944.3133948*

Yang, L., Sahli, H., Xia, X., Pei, E., Oveneke, M. C. and Jiang, D. (2017). Hybrid depression classification and estimation from audio video and text information, *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, Association for Computing Machinery, New York, NY, USA, p. 45–51.
URL: *https://doi.org/10.1145/3133944.3133950*