Received: May 15, 2024. Revised: July 9, 2024. 

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

# Correlation-Based Attention Weights Mechanism in Multimodal Depression Detection: A Two-Stage Approach

**Suresh Mamidisetti[1]*** **A. Mallikarjuna Reddy[2]**

[1]*Department of Computer Science and Engineering, Anurag University, Telangana, India*
[2]*Department of Artificial Intelligence (AI), Anurag University, Telangana, India*
* Corresponding author's Email: sureshmamidisetti@gmail.com

**Abstract:** Globally, the COVID-19 pandemic has aggravated the problem of depression. In the literature on depression detection, predictive capacity of multi-modal biomarkers like speech, facial and language characteristics were combined to improve the performance of Machine Learning (ML) models. The existing techniques have the limitation of not using the feature based weighting on multimodal data items in supervised learning models. The aim of the study is to use correlation based weighting mechanism between the feature and label in supervised learning to build an automatic multimodal depression detection system. We constructed a dataset of audio-visual recordings that consists of depressed and non-depressed subjects, to address this limitation by proposing a novel two-stage multimodal method for classification. The first stage attention weight is found using the Pearson correlation coefficients. Then these attention weights are combined with the multimodal feature vector, to use in the second stage. The resultant feature vector is experimented with three state-of-the-art classifiers i.e. Support Vector Machine (SVM), Decision Tree (DT) and K-Nearest Neighbour (KNN) classifiers, for predictions. Overall performance measures of proposed model are: accuracy 87.3%, precision 83.3%, recall 77%, and f1-score 80% which showed its effectiveness. Also outcomes of this proposed model were compared with the outcomes of existing state-of-the-art methods, such as baseline base classifiers and feature-level fusion method. Further, the proposed model is also evaluated applying to a widely used benchmarking dataset i.e. Distress Analysis Interview Corpus (DAIC) in depression detection. Results showed that combination of heterogeneous classifiers improves performance when classifying depressed and non-depressed subjects.

**Keywords:** Multimodal, Depression detection, Attention, Vocal-visual-verbal, Correlation, Fusion.

## 1. Introduction

Psychological disorders are significant obstacles to the deterioration of the global health agenda. Depression is one of the world's leading health concerns, and the COVID-19 pandemic is still exacerbating this problem [1]. Pandemic showed evident physical symptoms in individuals infected with the virus, but it also affected public mental health regardless of whether they are infected or not. Several factors like orders for home confinement (such as isolation, quarantine, and stay-at-home orders), lockdowns for indefinite periods are helpful to combat the virus, but these factors caused a surge in the incidence of depression during the Pandemic [2]. The World Health Organisation(WHO) also declared depression is an integral component of this pandemic [3]. Also during epidemics, the public is often fearful about getting infected with the disease, resulting in psychological ailments [4].

According to the American Psychiatric Association's (APA) Diagnostic and Statistical Manual of Mental Disorders-V (DSM-V), depression is a common mental disorder characterized by persistent feelings of sorrow and/or a significant lack of interest. Individuals suffering from depression may exhibit four or more of the following symptoms: weight changes, either loss or gain; disruptions in sleep patterns, including insomnia or hypersomnia; psychomotor retardation; fatigue or a significant loss of energy; impaired ability to think or concentrate;

feelings of worthlessness or excessive guilt; and suicidal ideation. These symptoms should persist at least for two weeks.

Methods and tools are the need of the hour in the health care sector to address the problem of depression. Recent advances in Machine Learning (ML) paved the way for the development of technologies that utilise biological indicators to detect a variety of mental, cognitive and neurological illnesses [5, 6]. ML based techniques are capable enough to analyse the real world data primitives for bringing technologies closer to the people.

Primarily, depression assessment methods are: i) structured interviews and ii) questionnaires. Health workers typically conduct structured interviews that depend on their skills, experience, and expertise. Questionnaires like Patient Health Questionnaire-9 (PHQ-9) have shown biases (responder reliability). Both these methods need the assistance of the health worker. When there is no availability of the expert or there is a stigma of the suspect to consult an expert for diagnosis, a tool/method for automatic depression diagnosis (ADD) is required, especially during the pandemic. ADD has various benefits such as it can be utilized as a pre-screening assessment for an individual at home before consulting with a clinician, it can help identify high-risk patients, reduce medical and travel expenses, and facilitate expedited diagnosis of depression and treatment. Early diagnosis of depression is critical to avoid its negative repercussions [7].

Earlier research revealed that Psychomotor Retardation (PR) is a central feature of depressive symptomatology [8]. PR apart from slowing down physical movements also impairs emotional behaviour. Generally, clinicians notice PR in the individual through direct behavioural observation of speech, facial expression, eye movements, speed, and degree of movements [9]. Hence, ideally, the ML based techniques can also use these behavioural cues that a clinician uses to classify depressed or non-depressed. To assess depression, we used emotion elicitation [10] and speech elicitation [11] activities to ascertain if the individual showed PR symptoms.

Emotion and speech elicitation are widely used activities in psychological assessment. In emotion elicitation activity, different kinds of photos/videos (like positive/neutral/negative) are shown to elicit the individual's emotions [12]. In speech elicitation activities individuals read a phonetically balanced paragraph and are given a spontaneous speech on the topic of their choice [13]. These emotion and speech elicitation activities are used to quantify the behavioural characteristics which are different to depressed and non-depressed subjects.

In the proposed work post data collection, multi modal features - vocal, visual and verbal indicators of depression are extracted from audio, video and text data sources respectively. This work presents novel two-stage architecture. In the first stage, two steps are carried out. First, attention weights are determined using correlation analysis based on association of features with their ground truth. In the next step these attention weights are utilised to quantify the importance of the features. In the second stage, these correlation based attention weights are applied on the multiple dataset settings by using different machine learning classifiers to predict depressed and non-depressed subjects.

The remaining sections of this paper are organised as follows: Background and related works are discussed in Section 2. Overview of the proposed method is described in detail in Section 3. The results obtained in the current study are presented and discussed in Section 4. And finally, Section 5 presents the conclusion.

## 2. Background and related works

In this section, first, we introduce depressive behaviour and then summarise different works on automatic depression detection.

### 2.1 Depressive behavior

Earlier studies support that depression and individual behaviour are closely associated [14]. Depressed individuals exhibit varied behaviour compared with non-depressed. Broadly, depressive behaviour can be categorized as verbal, visual, and vocal categories.

*Verbal behaviour:* early subjective research suggests that verbal communication may be a sign of depression in an individual [15-16]. For instance, depressed people tend to have fewer conversations and speak from a negative perspective [17].

*Visual behaviour:* earliest studies show that facial characteristics are critical cues of depression [18]. For example, depressed people tend to have fewer smiles, more frequent lip presses, shorter eye contact, reduced facial movements, etc. [19].

*Vocal behaviour:* studies found a strong association between paralinguistic or non-voiced cues and depression [20]. A few vocal characteristics of depressed people are lower pitch, longer pauses, less intensity, etc. [21].

All the aforementioned studies prove that the individual's verbal, visual, and vocal behaviours can be objective quantifiers of depression to build an automatic depression detection method.

Table 1. Overview of several works related to automatic depression detection

| Feature category A V T | Sample Features | Fusion Category E L H | ML Model | Dataset sources | Results | Ref erences |
|---|---|---|---|---|---|---|
| A T | Pitch, Negative valence word | No fusion | Random Forest | DAIC | MAE-5.84 RMSE-6.8 | [32] |
| A V | Formants, Pitch, HNR, left-right eye movment | E | SVM | Blackdog, Pitt, AVEC 2014 | Accuracy 74.6% | [33] |
| V | Eye movement features-gaze time, gaze shift and pupil size | No fusion | Kernel Extreme Learning Machine (KELM) | Real world dataset | Accuracy 76% | [34] |
| T | Unigram words, trigram words | No fusion | Logistic Regression | Twitter dataset | Accuracy 78% | [35] |
| T | Frequency of comments, number of followers, term frequency | No fusion | Multi-kernel SVM | SinaWeibo posts | Precision-75.56% f1-score-76.12% | [36] |
| T | Transcripts | No fusion | Attension mechanism | DAIC | MSE 3.78 | [37] |
| A | Pitch, jitter, shimmer | No fusion | Multiple linear regression model | CAPS (Psychosocial Care Centers) | $R^2$ was 0.508 & adjusted $R^2$ was 0.498 | [38] |
| A | eGeMAPS | No fusion | SVM and RF | mPower Voice Dataset | Accuracy-77% | [39] |
| A | Prosodic features | No fusion | LR | DAIC | Accuracy-68% | [40] |
| A | Pitch, Hormnics-to-noise ratio(HNR) | No fusion | SVM | PROMPT | Accuracy-75.3% | [41] |
| V T | First response time, eye activity and head pose | E L H | SVM | BDAS | Accuracy upto 88% | [7] |
| A V | Coverap, facial features | E | SVM | DAIC | Sensitivity upto 69% | [42] |
| A V T | Pitch, AU, eyegaze, headpose, negative word ratio | E | SVR and RF | DAIC | Root mean square error 4.02, Root MSE-5.09 | [25] |
| A V T | eGeMAPS, context features, AU, eyegaze, head pose | E | B-LSTM | DAIC | Root MSE-0.28 | [27] |
| A V T | Visual 2D features, Coverap features and transcripts | E | Multi level attension mechanism | DAIC | MAE-3.18 | [43] |
| A V T | Text CNN, audio CNN and video CNN | E | LSTM | DAIC | Accuracy-87% | [44] |

## 2.2 Automatic depression detection

Automatic depression detection using machine learning has two phases: training and testing the models. For this purpose, the dataset is created from individuals comprising of depressed and non-depressed. One of the most important aspects of training models is identifying predictive features to

classify between depressed and non-depressed. These features are extracted from the communication modalities involved in the dataset construction. Depending upon the number of communication modalities used in the automatic depression detection methods, they can be classified into three broad categories. They are unimodal methods, bimodal methods, and multi-modal methods, which are described here below.

*Unimodal methods:* These studies include works that use a single modality. In this method, features are selected from single data source either from verbal, visual, or vocal modality. For example, Caroline et al. in [22] recorded audio samples while the subjects (22 depressed and 11 control) interacted with the psychiatrist and used GNU Octave, to extract statistical features to train different ML classifiers. Wang et al. [23] recorded facial cues of depressed and healthy while showing positive/neutral/negative images. They extracted statistical features from various facial feature points to train the SVM classifier.

*Bimodal methods:* Studies in this method are formed with two modalities. Features are extracted from the combination of two data sources from verbal, visual, and vocal modalities. For instance, Alghowinem et al. [7] used audio-visual recordings during the interaction between the subject and clinician. They extracted variety of speech and facial features and then experimented with various fusion techniques to train the SVM classifier.

*Multi-modal methods:* Studies in this method incorporate inputs from three or more modalities [24]. For example, Guohou et al. [25] utilised features extracted from verbal, visual and vocal modalities to create a model.

Apart from methods based on communication channels, studies are generally fused with features of more than two or three modalities using early fusion. Early fusion is also known as feature fusion. Features extracted from the individual communication modalities are concatenated to form a new feature vector. Then, this resultant feature vector is fed into the ML classifier. Authors in [26, 27] fused audio, video, and text features to train a model.

Table 1 provides a brief overview of studies on automatic depression detection in various aspects such as feature category, sample features, fusion category (early, late or hybrid), ML techniques used, dataset sources, and reported results. Following observations are drawn from the table. They are: i) Studies did not focus on practical implications because the developed method could not be tested with a data sample of a new subject without clinician intervention. ii) In most studies, the datasets consist of multi-modal data items such as audio, video, text, etc. They did not give any kind of weights depending upon the strength of the modality. iii) Researchers did not examine the association of the features with the ground label and assign a weight as that the feature set becomes strong w.r.t modality used. Some strategy like if the modality is highly correlated higher weight or likely to assign lesser weight if the association is low, is not followed. iv) Due to dynamic psychological characteristics within the individual, there are variations in depressive behaviours. These variations cannot be ignored that overlook individual differences. Even though there are studies on multimodal cues, few studies ignore a modality, which may overlook the differences in the individual.

### 2.3 State of the art methods

Utilizing the multimodal cues like vocal, visual and verbal data sources to predict depression state using machine learning is point of the interest in the state-of-the-art works [28]. Few of the works are discussed here: Yang et al. [29] introduced the RLKT-MDD (Representation Learning and Knowledge Transfer for Multimodal Depression Diagnosis) model framework for depression classification problem. Their work utilised representation learning where model autonomously identifies significant patterns and identifies the features from diverse data sources. They employed DAIC dataset to significantly improve in the state-of-the-art works for depression classification. They achieved accuracy of about 78% utilising the representation learning and knowledge transfer framework. Junqi et al. [30] proposed multi-modal fusion model for depression detection that integrates multi-level audio features and text sentence embeddings. First, they extracted Low-Level Descriptors (LLDs), mel-spectrogram features, and wav2vec features from the audio data. Then they introduced a Multi-level Audio Features Interaction Module (MAFIM) to combine these features into a comprehensive audio representation. These models achieved performance of precision up to 70% and f1-score up to 68% on the DAIC dataset. Adefemi et al. [31] proposed an optimal algorithm for predicting depression severity by identifying personalized risk factors using machine learning. The potential benefits of their approach included enhanced accuracy in assessing severity, personalized treatment strategies, and improved identification of risk factors. Among the algorithms tested, the random forest (RF) algorithm proved to be the most effective,

demonstrating performance metrics of recall up to 73% on DAIC dataset.

## 3. Overview of the proposed work

Our present research aims to address the limitations discussed in section 2. By using a novel two-stage framework that leverages multimodal features extracted from audio-visual recordings collected during emotion and speech elicitation activities, we classified subjects into depressed and non-depressed.

### 3.1 Dataset acquisition

For data collection, volunteer participants were approached through personal contacts, social networks, and mailing lists. Among 225 responses obtained, 219 volunteers (~50% female, average age of 18 to 19 years) participated in our work. Participants were given appointments to participate in sequential sessions.

Data acquisition was performed in a well-lit room condition. The recording equipment's included a DELL C270 HD webcam for video, and a set of OPPO Enco airdopes with inbuilt microphone for audio. The resolution of the camera video is 720P, 30 fps. Each participant with airdopes was asked to sit on a chair in front of a windows OS installed laptop (placed on the table). A constant distance of approximately 1 metre (from eye to the laptop screen) was maintained throughout the process. A webcam was mounted on a laptop which aims the participant's upper body. The head of the participant and web cam were adjusted at the same height. The camera application runs in the background for audio-visual recording of participants. A video was played for the participant to conduct emotion and speech elicitation activities.

During the session, researchers conducted emotion and speech elicitation activities to record each subject's facial and speech responses. Table 2 lists the experimental procedure with time duration involved in emotion and speech elicitation activities. During the session, research assistants presented monitor with a wide screen/audio to perform emotion elicitation. The purpose of this activity was to elicit emotions through the presentation of different types of multimedia clips - positive, neutral, and sad taken from popular psychological film categories. The researchers explained participants that each video clip would be followed by a one-minute blank monitor, giving them time to erase every sensation, memories, and ideas from their minds. Following the screening of all the videos, participants given rest for ten minutes before starting speech elicitation exercise. During this exercise, participants were asked to provide a speech in two distinct formats. They were first assigned with reading aloud a story that had displayed on the monitor, namely a phonetically balanced paragraph called "The North and the South

Table 2. Emotion and speech elicitation are conducted through an experimental procedure with time duration

| Experimental tasks | Procedure | Description(source) | Duration (in Minutes:Secs) |
|---|---|---|---|
| Emotion Elicitation | Blank Monitor | NA | 1:00 |
| | Positive video | A highly rated freely available prank video in youtube that contains hilarious acts of comedy. | 4:10 |
| | Blank Monitor | NA | 1:00 |
| | Neutral video | Customised and self made video on shapes and neutral colours. | 3:20 |
| | Blank Monitor | NA | 1:00 |
| | Negative video | A sarrow scene when children and mother who is cancer patient were crying because of sudden death of their father in a road accident. | 4:20 |
| Break | Blank Monitor | NA | 10:00 |
| Speech Elicitation | Reading Passage | Short story called "the north and the south wind" | ~1:00 |
| | Free form speech | Free form of speech of their choice. | ~2:00 |

Wind." Next, they were asked to give a free form of speech of their choice. All the sessions were recorded for data pre-processing.

### 3.2 Ground truth labelling

The volunteers were given physical forms of PHQ-9 questionnaire at the data collection phase. On the basis of PHQ-9 scores, each volunteer's mental status was classified into binary classes (Non-depressed=0; those who exhibit indications of depression=1).

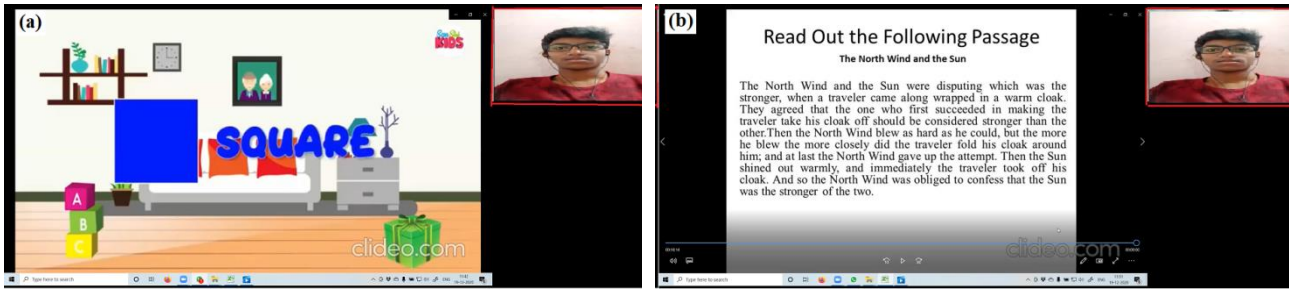
A psychologist was recruited on honorarium for ground truth labelling. The psychologist reviewed the PHQ-9 forms and conducted a one to one face individual session with the participant. To ensure the accuracy of the ground truth labels and build a more dependable dataset, this additional verification step was added. This technique produced a final dataset with labels for 137 non-depressed and 82 depressed people.

### 3.3 Feature extraction

The proposed method extracted multimodal features from visual, audio and verbal cues of the recorded audio-visual data. To begin with, the researcher retrieved visual cues from participants in the form of (.mp4) files, with an emphasis on facial reactions, as seen in the red boxes in Fig. 1. These clipped facial reactions were then processed further. Moving on to the recorded meeting during the speech elicitation phase, as shown in Fig. 1(b), speech cues in the form of (.mp3) files were obtained. To improve the quality of the recorded voice material, the SOX tool was used to remove the noise. Then, a research assistant transcribed the language spoke by each participant during the free-form speech. This transcription procedure used audio-visual recordings, with statements separated by at least 300 milliseconds of quietness.

Table 3. The list of selected features

| Category | Sub-category | Features | Description |
|---|---|---|---|
| Vocal | Prosodic | Pitch(va1), Intensity(va2), Pulses(va3), Amplitude(va4), Jitter(va5-va10), Shimmer(va11-va16), energy(va17), power(va18), formants(va19-va22) | These features represent fundamental acoustic characteristics of speech signal. (how they speak) |
| | Spectral | Hormonicity(va23-va25), autocorrelation(va26), MFCC(va27-va31), spectralflux(va32) | These features provide frequency domain parameters of speech biosignal. |
| | Non-voiced | Unvoicedframes(va33), voicebreaks(va34), Degreeofvoicebreaks(va35) | These features resemble the non-voiced measures of speech. |
| Visual | AUs | AU1(vb1), AU2(vb2), AU4-7(vb3-vb6), AU9-10(vb7-vb8), AU12(vb9), AU14-15(vb10-vb11), AU17(vb12), AU20(vb13), AU23(vb14), AU25(vb15), AU26(vb16), AU28(vb17), AU45(vb18) | These features represent any fundamental activity of movement on the face. |
| | Eye gaze | Righteyedirection(vb19-vb21), Lefteyedirection(vb22-vb24), Eyegazeangle(vb25-vb26) | These features provide the eye gaze activities (direction/angle of both eyes). |
| | Head Pose | Pitch(vb27), yaw(vb28), roll(vb29) | These features measure the head movement activities. |
| Verbal | Linguistic | #words(vc1), #types(vc2), TTR(vc3), Letters per word(vc4), #sentences(vc5), #of words per sentence(vc6), determiners(vc7), demonstratives(vc8), #pronouns, first_pronouns(vc9), second_pronouns(vc10), third_pronouns(vc11), conjuncts(vc12), connectives(vc13), negation(vc14), future(vc15) | These features resembles the lexical content and grammatical structure involved in their open form of speech (for ex. word choice, grammatical structure) |
| | Affect | #negative(vc16), #neutral(vc17), #positive(vc18), #compound(vc19), social(vc20), action(vc21), joy(vc22), ff(vc23), fd(vc24), f(vc25), wellbeing(vc26) | These components describe the semantic meaning involved in the spontaneous speech (what they speak) |




(a) (b)

Figure. 1 Samples during the use of elicitation techniques: (a) Emotion Elicitation: while watching the neutral video, participant's facial cues are recorded and (b) Speech Elicitation: participant's speech is recorded while they read the phonetically balanced paragraph

The three forms of modalities obtained from the previous process, were used to extract the multimodal features as indicated in Table 3. In the current study, vocal features were sub-categorised into prosodic, spectral, and non-voiced features based on the findings of studies on depression detection [37, 45-46]. The statistical features were calculated from the low level features obtained using Praat tool.

The physician uses visual descriptors during interaction with patients to diagnose depression [47]. Based on the literature on depression detection [33, 48-49], three kinds of low level visual features were extracted using OpenFace tool kit. They are facial Action Units (AU), eye gaze, and head pose features. Statistical feature vectors were then computed from these low-level features.

Researchers found a higher association in between verbal cues and depression [17]. Based on the findings of several earlier studies [50-53], we categorised verbal features into two sub-categories: linguistic and affect-related features. Linguistic features were extracted using the SiNLP software tool, and affect-related features were extracted using SEANCE tool.

### 3.4 Overview of the proposed two-stage architecture

The overview of the proposed Two-Stage architecture is shown in Fig. 2. The first stage in the proposed approach is to determine attention weights. In this step, Correlation-based attention weight is determined to focus on association of features with their ground truth. Mathematically, this attention mechanism can be represented as follows: Our Dataset(X) consists of all features, where each element X_i is associated with a Modality C_i. The correlation-based attention weight, denoted as A_i, for each element X_i can be computed as shown in Eq. (1):

$$A\_i = Correlation(X\_i, C\_i) \quad (1)$$

Here, Correlation represents a correlation function using Pearson correlation coefficients (elaborated in 3.4.1). The resulting attention weights A_i quantify the relevance of each element X_i with respect to its associated individual modality. Higher correlation values indicate stronger attention to that ground truth, while lower values imply reduced attention.

#### 3.4.1. Pearson's correlation coefficients

It is a statistical method to get an estimate of the linear relationship between two continuous variables X and Y. There are three assumptions that Pearson correlation coefficient uses. First, both of these variables need to be well-distributed. Second, there is a straight line relationship between the two variables, and third, this data is evenly distributed around the regression line. The Pearson correlation is given by the following formula shown in Eq. (2):

$$\rho X,Y = \frac{cov(X,Y)}{\sigma X\ \sigma Y} \quad (2)$$

Here cov(X,Y) represents covariance between X and Y and σX & σY indicates standard deviation of X and Y respectively.

The correlation is defined as a measure of linear relationship between two quantitative variables. It is also defined as a measure of how strongly one variable depends on another variable. The logic behind using correlation is that good variables are highly correlated with the target. Correlation between the target observations and the input features is very important. Higher the correlation then higher is the association. In that case the attention weight is preserved for the later stage. It is also used to find the strength of the linear relationship between two data

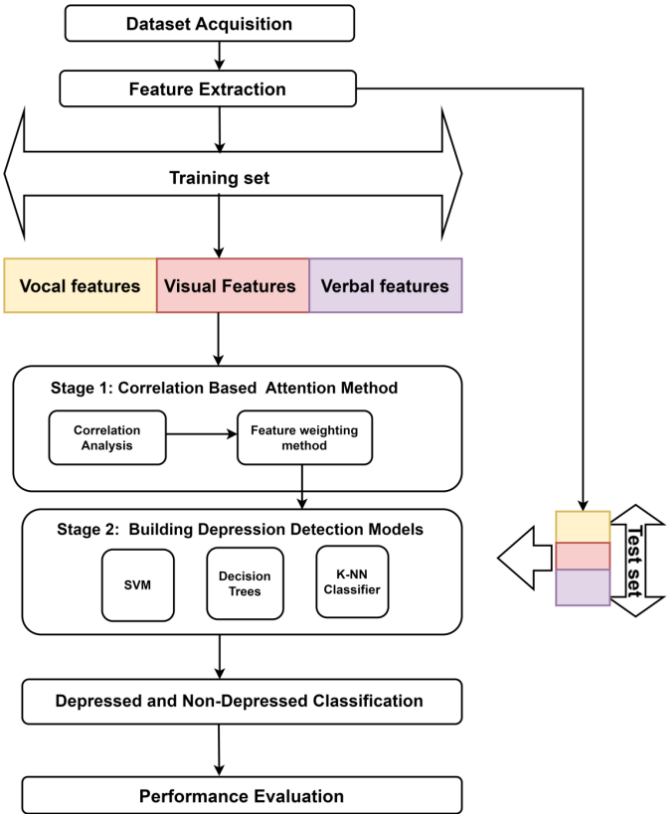



Figure. 2 Overview of the proposed two-stage architecture

variables, which can range from -1 to +1. -1 represents that there is a negative correlation, where the values of one variable increase as the values of the other decrease. 0 indicates that there is no linear correlation between the two variables. +1 implies a positive correlation, where the values of one variable increase as the values of the other increase.

The following Eq. (3) is used to calculate the value of the Pearson correlation coefficient-

$$A_{i_{xy}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y-\bar{y})^2}} \quad (3)$$

where,

n: Number of data points or observations in the dataset

$x_i$: Each individual data point in the variable X

$y_i$: Corresponding data point in the variable Y

$\bar{x}$: Mean of variable X

$\bar{y}$: Mean of variable Y

The attention score obtained in Eq. (1) is utilised in the following way: The feature weighting is computed with the attention scores (A_i), which quantify the importance of each feature (X_i) in the individual modality.

The weighted feature vector (X_w) is represented mathematically as shown in Eq. (4):

$$X_w = (A_1 \times X[1], A_2 \times X[2], \ldots., A_n \times X[n]) \quad (4)$$

Here, X[1], X[2], ..., X[n] denotes the values of the individual features in our dataset. While A_i signifies the attention score for feature X_i that represents the importance of the feature. Higher the value higher the importance. These attention scores (A_i) are determined through statistical analysis that builds final feature vector.

In theory, the proposed work can lead to effective classification for the following reasons: i) Features with higher attention scores are given greater influence in the model's decision-making process, allowing it to focus on the most relevant aspects of the dataset. ii) Feature vector is built by attention scores that contribute to building a robust feature vector that reflects the significance of individual modality. This robustness in feature representation helps in mitigating the influence of noise or less relevant features, thus contributing to better generalization and improving model performance. iii) Pearson correlation coefficient considers multivariate pair wise relationships between features and ground truths in individual modality, providing a multivariate perspective. By integrating these associations into feature weights, this method

accounts for the collective impact of features, enhancing its ability to capture complex patterns in the dataset. iv) Attention feature weighting can imply an implicit form of dimensionality reduction by assigning lower weights to less influential features. This can help to a simplified and more interpretable representation of the dataset, potentially improving model interpretability.

### 3.5 Building depression detection model

In the current work, we proposed a novel set of multimodal feature vector and two-stage framework for depression classification. The first stage has three sub-models, each of which corresponds to a different modality. Each sub-model takes input from individual modality features (i.e. vocal, visual and verbal) and outputs values for the classification of the depressed or non-depressed subjects with the attention scores obtained in stage one. Following experimentation with the associated vocal, visual, and verbal features using three state-of-the-art classifiers - SVM, DT, and KNN respectively. 80 percent of sample size was used as training set and 20 percent of sample size was used as test set.

## 4. Results

### 4.1 Experimental procedure

To form multimodal features, audio, video, and transcripts were fed into the Praat, OpenFace and SEANCE toolkits respectively. These are state-of-the-art (SOTA) tools for low level feature extraction in their respective domains. Statistical features (like mean, count) are computed and then normalised. Next, we conducted experiments on the proposed two-stage approach. In the first stage, we determined the attention weights associated with the individual modality features and ground truth.

In the second stage, we experimented with three state-of-the-art classifiers, SVM, DT, and KNN. In addition, we evaluated the proposed model performance in terms of various performance metrics. Furthermore, we also trained and tested several baseline models like using individual modality features and feature level concatenation methods. We evaluated the proposed approach's effectiveness in depression detection to that of existing baseline models. Finally, we used a depression detection benchmark dataset called the DAIC to validate the performance of our proposed approach.

### 4.2 Performance evaluation

The effectiveness of any classification model can be measured in terms of performance metrics. As shown in Table 4, classification results are computed using confusion matrix.

Table 5 illustrates the performance metrics used in the current work. Accuracy, precision, recall, f1-score, and Mathew's correlation coefficient (MCC).

Table 4. Confusion matrix

| **Confusion matrix** | **Positive(actual) (P=TP+FN)** | **Negative(actual) (N=TN+FP)** |
|---|---|---|
| Positive (Test outcome) | True Positive(TP) | False Positive(FP) |
| Negative (Test outcome) | False Negative(FN) | True Negative(TN) |

Table 5. Performance metrics used in the current work

| **Performance Metric** | **Formula** | **Description** |
|---|---|---|
| Accuracy(acc) | (TP+TN)/(TP+FP+TN+FN) | It is ratio of correct predictions over total number of observations assessed. |
| Precision(p) | TP/(TP+FP) | It is the ratio between true positive and all the positives. |
| Recall(r) | TP/(TP+FN) | It is the ratio of positive observations that are correctly classified. |
| f1-score(F) | 2.p.r/(p+r) | It is harmonic mean of precision and recall. |
| Mathew's correlation coefficient(mcc) | (TP × TN - FP × FN)/sqrt((TP+ FP) × (TP+FN) × (TN+FP) × (TN+FN)) | It is used to evaluate the binary classifier, for diagnosis of depression in a patient. |

### 4.3 Performance evaluation settings

To evaluate the performance of the proposed work, we opted two methods: In the first method, performance measures are compared to models that use unimodal, bimodal, and multimodal features to determine the impact of multimodal features on performance over unimodal and bimodal features. State-of-the-art classifier in depression classification i.e. Logistic Regression (LR) is selected to train (80%) and test (20%) a model using unimodal, bimodal and multimodal features [7]. The performances of our selected State-of-the-art classifiers (SVM, DT, and KNN) are referred as P_SVM, P_DT, and P_KNN respectively. In the second method, we chose two types of baseline models to assess the impact of the proposed approach in depression detection. In the first baseline method, Random Forest (RF), Extra Trees (EX) and Naive Bayes (NB) classifiers were used with the concatenated multimodal feature vector to train (75%) and test (25%) a model. These classifiers were used because they performed well in the individual-level modality. B1_RF, B1_EX, B1_NB refers to first baseline models performances using RF, EX and NB classifiers respectively. Similarly, in the second baseline model, Decision Trees (DT) was used to create a model with multimodal features. B2_D indicates second baseline model performance using DT classifier. We have selected Decision Tree classifier because it has been widely used from decades to recent works in medical diagnosis problems [26, 54]. It is also proved that Decision Trees are efficient in depression classification problems using multimodal data items [55].

### 4.4 Depression detection results

Table 6 shows the comparison of the performance metrics with different feature settings using LR classifier. The purpose of this comparison is to gain insights into the impact of multimodal features over unimodal and bimodal features. From Table 6, it is evident that combination of features from different modalities leads to improvement of the models performance metrics. It is also evident that models using bimodal features gave better performance measures than unimodal features. Similarly multimodal feature vector gave better results than bimodal feature vectors. It is evident that our findings showed that combination of features lead to improved performance of the model.

Table 6. Performace of multimodel features with Logistic Regression classifier

| Input features | Performance metrics | | | | |
|---|---|---|---|---|---|
| | acc | p | r | F | mcc |
| Vo | 54.1 | 64.2 | 60.0 | 62.0 | 0.043 |
| Vi | 58.3 | 63.6 | 53.8 | 58.3 | 0.174 |
| Ve | 54.1 | 58.3 | 53.8 | 56.0 | 0.083 |
| Vo+Vi | 62.8 | 71.4 | 66.6 | 68.9 | 0.218 |
| Vi+Ve | 62.5 | 64.2 | 69.2 | 66.6 | 0.240 |
| Vo+Ve | 62.5 | 58.3 | 63.6 | 60.8 | 0.250 |
| Vo+Vi+Ve | 70.0 | 66.0 | 72.7 | 69.5 | 0.418 |

Table 6 reports that visual features gave better than vocal and verbal features in terms of accuracy in unimodal feature settings. Model comprising of visual features gave better performance measures than combining verbal and vocal features in terms of all performance metrics in bimodal feature settings. These findings indicate that visual features have more importance than vocal and verbal markers in depression detection. Here Vo, Vi and Ve represent the vocal, visual and verbal features respectively.

Table 7 shows that our proposed two-stage framework model outperforms the baseline methods across all the performance metrics. Therefore, these findings provide strong support for the importance of two-stage framework for depression classification. It is further observed that model with P_SVM performed better in terms of accuracy (87.3%), precision (83.3%) and f1-score (80.0%) than with P_DT and P_KNN. P_SVM outperformed other two baseline methods with accuracy of 87.3%, precision of 83.3%, recall of 76.9%, f1-score of 80.0%, and

Table 7. Performance of proposed method over baseline methods

| Performance metric | P_SVM | P_DT | P_KNN | B1_RF | B1_EX | B1_NB | B2_D |
|---|---|---|---|---|---|---|---|
| acc | 87.3 | 79.1 | 78.2 | 70.8 | 62.5 | 62.5 | 75.0 |
| p | 83.3 | 75.0 | 69.3 | 69.2 | 61.5 | 58.3 | 72.9 |
| r | 76.9 | 81.8 | 80.2 | 75.0 | 66.6 | 63.4 | 72.7 |
| f | 80.0 | 78.2 | 75.0 | 72.2 | 64.0 | 60.8 | 72.7 |
| mcc | 0.585 | 0.585 | 0.485 | 0.418 | 0.250 | 0.250 | 0.496 |

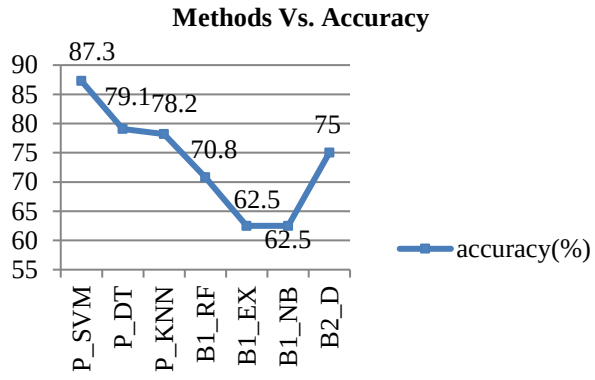



Figure. 3 Comparison of methods in terms of accuracies

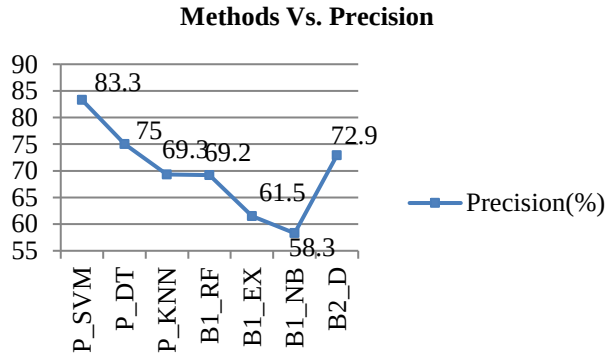



Figure. 4 Comparison of methods w.r.t precision values

mcc of 0.585. These findings suggest that it is important to determine the weight of the feature vector in supervised learning environment.

Fig. 3 shows a line graph of accuracy metric with the proposed approach performances of SVM, DT and KNN against two baseline methods. From the graph it can be observed that SVM achieved higher accuracy over DT, KNN and baseline methods, scoring 87.3%. This indicates that the proposed method, especially when implemented with SVM, superior results compared to the baseline methods, showcased its efficacy in enhancing classification accuracy for the given task. By assigning the higher attention scores to feature vector with stronger linear correlations, the technique effectively prioritized relevant information for the classification task. This has reduced the impact of noise and redundancy but also optimized the SVM model's learning process, allowing it to create a more accurate decision boundary. The attention based methods determination of multivariate relationships and its alignment with SVM characteristics further contributed to enhance feature discrimination and improved overall accuracy, surpassing the performance of baseline methods.

Fig. 4 shows a line graph of precision metric with the proposed approach performances of SVM, DT and KNN over two baseline methods. From the graph it is can be seen that SVM achieved higher precision over DT, KNN, and baseline methods scoring 83.3%. Precision is a crucial metric, particularly in situations where minimizing false positives is essential. The higher precision recorded by SVM indicates its effectiveness in correctly identifying relevant instances and reducing the likelihood of mis-classifications. This result suggests that the proposed method, especially when accomplished using SVM, exhibits a superior ability to provide precise and accurate classifications.

Fig. 5 shows the recall scores of the proposed approach performances of SVM, DT and KNN over two baseline methods. Here it can be seen that DT recorded higher performance than SVM, KNN and baseline methods. DT reported 81.8% of recall value which is higher than all others in the current work. Recall is a critical metric in situation where the identification of all relevant instances is crucial. The higher recall attained by DT indicates its performance

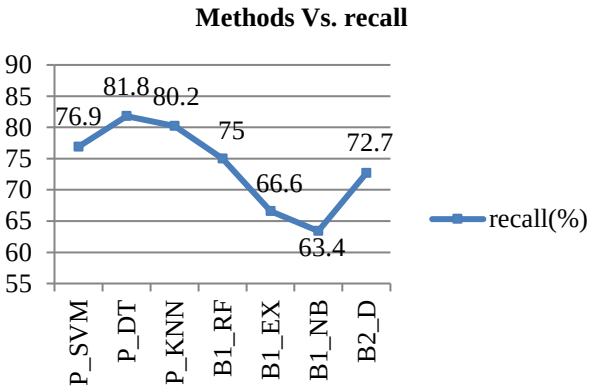



Figure. 5 Comparison of methods in terms of recall values

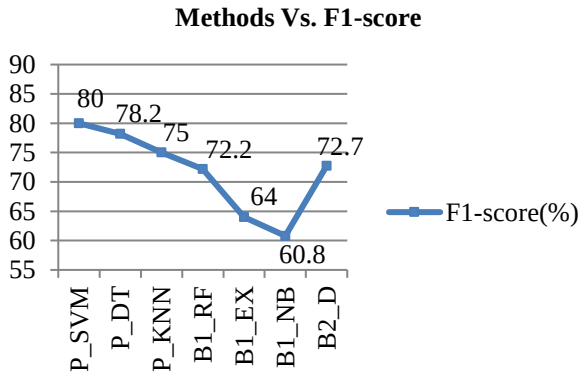



Figure. 6 Comparison of methods in terms of f1-score values

in capturing a larger proportion of positive instances, minimizing false negatives. This result supports that, within the context of the current work, the DT method reports in identifying and recalling instances of interest more effectively. The specific circumstances of the dataset or the nature of the problem have favoured the Decision Tree's ability to capture true positives, contributing to its superior recall performance.

Fig. 6 shows the f1-scores of proposed approach SVM, DT and KNN over two baseline methods. Here it can be seen that SVM recorded higher performance than DT, KNN and baseline methods. SVM reported 80.0% of f1-score value which is higher than all others in the current work. The f1-score is a balanced metric that considers both precision and recall, making it suitable for this situation where achieving a balance between false positives and false negatives is crucial. The higher f1-score supports by SVM shows its effectiveness in achieving a supportive blend of precision and recall, thereby offering a robust performance in the current work. This result suggests that the proposed work, particularly when integrated with SVM, excels in striking a balance between accurate positive classifications and comprehensive comparison of relevant instances, outperforming all other methods in the current situation of f1-score.

### 4.5 Validation of proposed method with benchmarking dataset

To illustrate the performance of our proposed two-stage framework, it is evaluated on publicly accessible benchmarking dataset in the depression classification studies called DAIC dataset [56-57]. This data set is a multimodal collection of semi-structured clinical interviews conducted by an animated virtual interviewer with a participant. The data collected includes audio, video, transcripts of the interviews designed to support diagnosis of depression. Binary labels(depressed/non-depressed) were provided for all the 188 subjects using standardised PHQ-8 questionnaire [58].

Table 8 shows the comparison of the performance metrics between the proposed method and SOTA approaches using DAIC dataset. Observations from table 8 provide the following findings: This analysis compares various methodologies and performance metrics for depression classification using multimodal approaches. The methodologies involve different machine learning techniques, and the performance metrics include accuracy, precision, recall, and f1-score. These Automatic depression detection methods utilized DAIC dataset that included audio, video, and semantic elements. Hence these methods are chosen to facilitate the comparison of the effectiveness of the proposed model. The SOTA techniques that are introduced in the introduction section are utilised for comparison. Yang et al. in [29] employed representation learning and knowledge transfer techniques to improve the diagnosis of depression. These methods leverage multimodal data, potentially combining text, audio, and visual inputs but lacks the weighting mechanism. It can be stated that our approach outperformed their approach by 1% because of the attention weights mechanism that gave importance to the individual modality. Junqi et al. in [30] integrated multi-level audio features and text sentence embeddings using a multi-modal fusion model. Our proposed approach outperformed precision by 4% and f1-score by 4%. From these findings it can be supported that utilising vocal, visual and verbal indicators of the depression

Table 8. Comparison of the performance metrics between the proposed method and SOTA approaches using DAIC dataset

| S.No | Authors /year | Methodology | Performance metrics |
|---|---|---|---|
| 1 | Yang et al./2024 | Representation learning and knowledge transfer for multimodal depression diagnosis | Accuracy of 78% |
| 2 | Junqi et al./2024 | Multi-modal fusion model that integrates multi-level audio features and text sentence embeddings | Precision upto 70% and f1-score 68% |
| 3 | Adefemi et al./2024 | Optimal algorithm test based on seven classifiers | Recall upto 73% |
| 4 | Our work | Our two-stage approach using SVM, DT and KNN classifiers | Accuracy 79.1%, Precision 74.2%, Recall 73.9%, f1-score 72.0% |

helped the machine learning techniques to distinguish depressed and non-depressed subjects. Adefemi et al. in [31] utilised seven traditional classifiers on multimodal depression indicators. From these results it can be indicated that the proposed attention weighting mechanism gained performance to outperform the recall metric by 1% that demonstrated the effectiveness of the current work.

## 5. Conclusion

Depression offers a unique context for medical diagnosis problems because symptoms are multi-dimensional. Machine Learning in health care applications has taken advantage of technological advances to measure signs of depression from multiple modalities. In the current work, we utilised combined predictive capacity of multi-modal biomarkers like spoken speech, facial and language characteristics which are different in depressed and non-depressed individuals to develop ML models. Here we collected data by performing popular psychological methods known as emotion and speech elicitation. This paper proposes a novel two-stage multimodal method for classifying depressed and non-depressed subjects. The first stage attention weights were found using the statistical method of Pearson correlation coefficients. These attention weights were found on the training data and used on the test data. Next stage uses the first stage's attention weights to combine with the multimodal feature vector. The resultant feature vector is experimented with three state-of-the-art classifiers i.e. Support Vector Machine (SVM), Decision Tree (DT) and K-Nearest Neighbour (KNN) for predictions. The overall performance measures of proposed model are: accuracy 87.3%, precision 83.3%, recall 77%, and f1-score 80%, which show the model's effectiveness. Also outcomes are compared with existing state-of-the-art methods such as baseline base classifiers and feature-level fusion method. Further, the proposed model was evaluated when applied to a widely used benchmarking dataset in depression detection (DAIC). Our findings conclude that combination of multimodal features with attention weighting feature strategy gains performance in classifying depressed and non-depressed subjects.

## Conflicts of Interest

The authors declare no conflict of interest.

## Author Contributions

Conceptualization, methodology, software and original draft preparation are done by Suresh Mamidisetti; supervision, review, and formal analysis are done by A. Mallikarjuna Reddy.

## References

[1] WHO-World Health Organization, "The impact of COVID-19 on mental, neurological and substance use services", 2020. https://www.who.int/publications/i/item/9789240012455 (accessed 07 November 2023).

[2] J. Bueno-Notivol, P. Gracia-García, B. Olaya, I. Lasheras, R. López-Antón, and J. Santabárbara, "Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies", *International Journal of Clinical and Health Psychology*, Vol. 21, No. 1, 100196, 2021.

[3] WHO, "Mental health preparedness and response for the COVID-19 pandemic: report by the Director-General", 2021. https://apps.who.int/iris/handle/10665/359717 (accessed 07 November 2023).

[4] R. C. W. Hall, and M. J. Chapman, "The 1995 Kikwit Ebola outbreak: lessons hospitals and physicians can apply to future viral epidemics", *General Hospital Psychiatry*, Vol. 30, No. 5, pp. 446-452, 2008.

[5] R. Strawbridge *et al*., "Care pathways for people with major depressive disorder: A European Brain Council Value of Treatment study", *European Psychiatry*, Vol. 65, No. 1, 2022.

[6] N. A. Dewan, J. S. Luo, and N. M. Lorenzi, *Mental Health Practice in a Digital World*, Springer Cham, Switzerland, 2015.

[7] S. Alghowinem *et al*., "Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors", *IEEE Transactions on Affective Computing*, Vol. 9, No. 4, pp. 478-490, 2018.

[8] N. Dantchev, and D. J. Widlöcher, "The measurement of retardation in depression", *Journal of Clinical Psychiatry*, Vol. 59, No. 14, pp. 19-25, 1998.

[9] Z. S. Shah, K. Sidorov, and D. Marshall, "Psychomotor cues for depression screening", In: *Proc. of International Conf. On Digital Signal Processing (DSP)*, London, UK, 2017.

[10] J. Coan and J. J. B. Allen, *The Handbook of Emotion Elicitation and Assessment*, Oxford University Press, United Kingdom, Vol. 46 2008.

[11] P. A. Barbosa, and S. Madureira, "Elicitation techniques for cross-linguistic research on professional and non-professional speaking styles", In: *Proc. of International Conf. On Speech Prosody*, Boston, USA, pp. 503-507, 2016.

[12] J. J. Gross and R. W. Levenson, "Emotion Elicitation using Films", *Cognition and Emotion*, Vol. 9, No. 1, pp. 87-108, 1995.

[13] B. Stasak, J. Epps, and R. Goecke, "Elicitation design for acoustic depression classification: An investigation of articulation effort, linguistic complexity, and word affect", In: *Proc. of International Conf. On Interspeech*, Stockholm, Sweden, pp. 834-838, 2017.

[14] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, *Multimodal Assessment of Depression from Behavioral Signals*, ACM, New York, Vol. 2, 2018.

[15] S. Newman, and V. G. Mather, "Analysis of spoken languauge of patients with affective disorders", *American Journal of Psychiatry*, Vol. 9, No. 4, pp. 912-942, 1938.

[16] G. A. Miller, *Language and Communication*, McGraw-Hill, New York, 1963.

[17] B. Stasak, "An Investigation of Acoustic, Linguistic, and Affect Based Methods for Speech Depression Assessment", (Doctoral dissertation), UNSW, Sydney, 2018. http://hdl.handle.net/1959.4/61278.

[18] P. H. Waxer, "Therapist training in nonverbal communication I: Nonverbal cues for depression", *Journal of Clinical Psychology*, Vol. 30, No. 2, pp. 215-218, 1974.

[19] A. Pampouchidou *et al*., "Automatic Assessment of Depression Based on Visual Cues: A Systematic Review", *IEEE Transactions of Affective Computing*, Vol. 10, No. 4, pp. 445-470, 2019.

[20] P. F. Ostwald, "Acoustic Methods in Psychiatry", *Scientific American*, Vol. 212, No. 3, pp. 82-91, 1965.

[21] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis", *Speech Communication*, Vol. 71, pp. 10-49, 2015.

[22] C. W. Espinola, J. C. Gomes, J. M. S. Pereira, and W. P. dos Santos, "Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study", *Research on Biomedical Engineering*, Vol. 37, No. 1, pp. 53-64, 2021.

[23] Q. Wang, H. Yang, and Y. Yu, "Facial expression video analysis for depression detection in Chinese patients", *Journal of Visual Communication and Image Representation*, Vol. 57, pp. 228-233, 2018.

[24] M. Morales, S. Scherer, and R. Levitan, "A Cross-modal Review of Indicators for Depression Detection Systems", In: *Proc. of the Fourth Workshop on Computational Linguistics and Clinical Psychology*, Vancouver, Canada, pp. 1-12, 2017.

[25] S. Guohou, Z. Lina, and Z. Dongsong, "What reveals about depression level? The role of multimodal features at the level of interview questions", *Information & Management*, Vol. 57, No. 7, 2020.

[26] A. Pampouchidou *et al.*, "Depression assessment by fusing high and low level features from audio, video, and text", In: *Proc. of Int. Workshop On Audio/Visual Emotion Challenge*, ACM, New York, pp. 27-34, 2016.

[27] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, R. Garg, "Multi-level Attention network using text, audio and video for Depression Prediction", In: *Proc. of International Audio/Visual Emotion Challenge and Workshop (AVEC-19)*, Nice, France, 2019.

[28] Govind, G. Ansari, A. Sharma, P. Arya, and Y. Saxena, "Multimodal Depression Detection System Using Machine Learning", In: *Proc. of Int. Conf. On Informatics (ICI)*, Noida, India, pp. 1-7, 2024.

[29] S. Yang, L. Cui, L. Wang, T. Wang, and J. You, "Enhancing multimodal depression diagnosis through representation learning and knowledge transfer", *Heliyon*, Vol. 10, No. 4, e25959, 2024.

[30] J. Xue, R. Qin, X. Zhou, H. Liu, M. Zhang, and Z. Zhang, "Fusing Multi-Level Features from Audio and Contextual Sentence Embedding from Text for Interview-Based Depression Detection", In: *Proc. of Int. Conf. On Acoustics, Speech and Signal Procesing*, Seoul, Korea, pp. 6790-6794, 2024.

[31] A. Ayodele, A. Adetunla, and E. Akinlabi, "Prediction of Depression Severity and Personalised Risk Factors Using Machine Learning on Multimodal Data", *International journal of online and biomedical engineering*, Vol. 20, No. 09, pp. 130-143, 2024.

[32] L. Zhang, J. Driscol, X. Chen, and R. H. Ghomi, "Evaluating Acoustic and Linguistic Features of Detecting Depression Sub-Challenge Dataset", In: *Proc. of International Audio/Visual Emotion Challenge and Workshop (AVEC-19)*, Nice, France, pp. 47-53, 2019.

[33] S. M. Alghowinem, T. Gedeon, R. Goecke, J. Cohn, and G. Parker, "Interpretation of

depression detection models via feature selection methods", *IEEE Transactions on Affective Computing*, Vol. 14, No. 1, pp. 133-152, 2023.

[34] M. Li *et al.*, "Method of Depression Classification Based on Behavioral and Physiological Signals of Eye Movement", *Complexity*, 4174857, 9 pages, 2020.

[35] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts", *Computers in Biology and Medicine*, Vol. 135, pp. 104499, 2021.

[36] Z. Peng, Q. Hu, and J. Dang, "Multi-kernel SVM based depression recognition using social media data", *International Journal of Machine Learning and Cybernetics*, Vol. 10, No. 1, pp. 43-57, 2019.

[37] K. Milintsevich, K. Sirts, and G. Dias, "Towards automatic text-based estimation of depression through symptom prediction", *Brain Informatics*, Vol. 10, No. 1, 2023.

[38] W. J. Silva, L. Lopes, M. K. C. Galdino, and A. A. Almeida, "Voice Acoustic Parameters as Predictors of Depression", *Journal of Voice*, Vol. 38, No. 1, pp. 77-85, 2021.

[39] Y. Ozkanca, M. G. Öztürk, M. N. Ekmekci, D. C. Atkins, C. Demiroglu, and R. H. Ghomi, "Depression screening from voice samples of patients affected by Parkinson's disease", *Digital Biomarkers*, Vol. 3, No. 2, pp. 72-82, 2019.

[40] P. Wu, R. Wang, H. Lin, F. Zhang, J. Tu, and M. Sun, "Automatic depression recognition by intelligent speech signal processing: A systematic survey", *CAAI Transactions on Intelligence Technology*, Vol. 8, No. 3, pp. 701-711, 2023.

[41] B. Sumali *et al.*, "Speech quality feature analysis for classification of depression and dementia patients", *Sensors*, Vol. 20, No. 12, pp. 1-17, 2020.

[42] S. Yasin, A. Othmani, I. Raza, and S. A. Hussain, "Machine learning based approaches for clinical and non-clinical depression recognition and depression relapse prediction using audiovisual and EEG modalities: A comprehensive review", *Computers in Biology and Medicine*, Vol. 159, p. 106741, 2023.

[43] M. Fang, S. Peng, Y. Liang, C. C. Hung, and S. Liu, "A multimodal fusion model with multi-level attention mechanism for depression detection", *Biomedical Signal Processing and Control*, Vol. 82, p. 104561, 2023.

[44] Vandana, N. Marriwala, and D. Chaudhary, "A hybrid model for depression detection using deep learning", *Measurement: Sensors*, Vol. 25, p. 100587, 2023.

[45] M. Yamamoto *et al.*, "Using speech recognition technology to investigate the association between timing-related speech features and depression severity", *PLoS One*, Vol. 15, No. 9 September, pp. 1-10, 2020.

[46] Z. Liu, H. Kang, L. Feng, and L. Zhang, "Speech Pause Time: A Potential Biomarker for Depression Detection," In: *Proc. of International Conf. on Bioinformatics and Biomedicine*, Kansas City, USA, 2017.

[47] A. Pampouchidou (2018), *Automatic detection of visual cues associated to depression* (Doctoral Thesis), UBFC, France.

[48] M. Gavrilescu and N. Vizireanu, "Predicting depression, anxiety, and stress levels from videos using the facial action coding system", *Sensors*, Vol. 19, No. 17, 2019.

[49] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear, "Head pose and movement analysis as an indicator of depression", In: *Proc. of International Conf. on Affective Computing and Intelligent Interaction*, Geneva, Switzerland, pp. 283-288, 2013.

[50] D. Smirnova, P. Cumming, E. Sloeva, N. Kuvshinova, D. Romanov, and G. Nosachev, "Language patterns discriminate mild depression from normal sadness and euthymic state", *Frontiers in Psychiatry*, Vol. 9, 2018.

[51] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in reddit social media forum", *IEEE Access*, Vol. 7, pp. 44883-44893, 2019.

[52] K. Uhl, L. F. Halpern, C. Tam, J. K. Fox, and J. L. Ryan, "Relations of Emotion Regulation, Negative and Positive Affect to Anxiety and Depression in Middle Childhood", *Journal of Child and Family Studies*, Vol. 28, No. 11, pp. 2988-2999, 2019.

[53] B. Stasak, J. Epps, and R. Goecke, "Automatic depression classification based on affective read sentences: Opportunities for text-dependent analysis", *Speech Communication*, Vol. 115, pp. 1-14, 2019.

[54] J. R. Quinlan, "Induction of decision trees", *Machine Learning*, Vol. 1, No. 1, pp. 81-106, 1986.

[55] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision tree based depression classification from audio video and language information", In: *Proc. of International Workshop on Audio/Visual Emotion Challenge*

*(AVEC)*, Amsterdam, Netherlands, pp. 89-96, 2016.

[56] F. Ringeval *et al.*, "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition", In: *Proc. of International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Nice, France, 2019.

[57] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews", In: *Proc. of International Conf. on Language Resources and Evaluation (LREC)*, pp. 3123-3128, 2014.

[58] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population", *Journal of Affective Disorders*, Vol. 114, No. 1-3, pp. 163-173, 2009.