Multimedia Tools and Applications (2023) 82:4787–4820
https://doi.org/10.1007/s11042-022-12315-2

1215: MULTIMODAL INTERACTION AND IOT APPLICATIONS

A novel multi-modal depression detection approach based on mobile crowd sensing and task-based mechanisms

Ravi Prasad Thati[1] · Abhishek Singh Dhadwal[1] · Praveen Kumar[1] · Sainaba P[2]

Received: 31 March 2021 / Revised: 20 September 2021 / Accepted: 17 January 2022 /
Published online: 11 April 2022

Abstract
Depression has become a global concern, and COVID-19 also has caused a big surge in its incidence. Broadly, there are two primary methods of detecting depression: Task-based and Mobile Crowd Sensing (MCS) based methods. These two approaches, when integrated, can complement each other. This paper proposes a novel approach for depression detection that combines real-time MCS and task-based mechanisms. We aim to design an end-to-end machine learning pipeline, which involves multimodal data collection, feature extraction, feature selection, fusion, and classification to distinguish between depressed and non-depressed subjects. For this purpose, we created a real-world dataset of depressed and non-depressed subjects. We experimented with: various features from multi-modalities, feature selection techniques, fused features, and machine learning classifiers such as Logistic Regression, Support Vector Machines (SVM), etc. for classification. Our findings suggest that combining features from multiple modalities perform better than any single data modality, and the best classification accuracy is achieved when features from all three data modalities are fused. Feature selection method based on Pearson's correlation coefficients improved the accuracy in comparison with other methods. Also, SVM yielded the best accuracy of 86%. Our proposed approach was also applied on benchmarking dataset, and results demonstrated that the multimodal approach is advantageous in performance with state-of-the-art depression recognition techniques.

Keywords Depression detection · Multi-modal · Mobile crowd sensing · Emotion elicitation · Speech elicitation · Machine learning

1 Introduction

Depression has been a worldwide concern for a long time and continues to plague the global health agenda. According to the World Health Organization (WHO), more than 350 million

✉ Ravi Prasad Thati
thati.raviprasad@gmail.com

[1] Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology, South Ambazari Road, Nagpur, 440010, Maharashtra, India

[2] Department of Applied Psychology, Central University of Tamil Nadu, Tamilnadu, India

individuals are estimated to suffer from depression. It is equivalent to 4.4% of the world's population. Depression is forecasted to become the world's leading health concern by the year 2030 [43]. COVID-19 has forced people throughout the world to stay indoors and minimize social interactions, Thus exacerbating the depression situation [32]. During the pandemic, the prevalence of depression in the general population is estimated to be 33%. COVID-19 not only impacts physical health concerns but also results in several mental illnesses. However, early diagnosis followed by appropriate treatment has proven to be successful in reducing its impact. Therefore, methods and tools for monitoring mental health are an immediate requirement [53].

Traditional methodologies rely on self-report or clinician consultation for spotting mental illness of an individual. Self-report has known limitations like inadvertence while filling up questionnaires and may be deemed unreliable. Clinician consultation depends on the physician's expertise, patient's budget, doctor's availability, and various other parameters. Hence, auxiliary methodologies for detection of psychological ailments are essential and have drawn the attention of researchers to assist in early diagnosis of depression [26].

Several non-traditional strategies exist for depression detection. Ideally, all the indicators that a clinician utilizes could be modelled into machine learning(ML) algorithms to diagnose depression [38]. The majority of the approaches rely on various characteristics like visual manifestations, acoustic and linguistic communication, smartphone usage activities, social media content, physiological cues, etc. [14, 21, 24, 34, 45, 47, 48, 56]. Few approaches combine different modalities like visual with speech [38, 42].

Combining different modalities is not trivial as each modality behaves differently. For example, each modality source is in a different form (images in visual domain, and text in social media content). From the available literature in the field, it is evident that integrating the modalities provides promising results [4, 42, 46].

Collaborative efforts of researchers across various fields are involved in depression detection. Each subfield has different data sources with slightly different goals. Broadly, the majority of the data collection is based on Task/interview-based and mobile crowd sensing-based procedures. Many Task/interview-based depression detection methods consider only in-situ representations such as facial cues, pitch of the voice, etc. Several mobile crowd sensing based depression detection methods take into real-time (days/nights) markers from smartphones like location information, accelerometer, etc. These approaches are inadequate to address the problem completely as Task-based approaches lack historical context. In contrast, mobile crowd sensing-based approaches ignore non-verbal and verbal indicators that clinician primarily rely on. Thus, a methodology that addresses these limitations by combining these approaches is the need of the hour.

With the advent of technology, smart phone has truly become an essential service for an individual in today's world. Mobile Crowd Sensing(MCS) refers to a wide range of methods using mobile devices capable of sensing and computing, in which people share data and derive information to quantify and record behaviours of mutual interest. MCS is a key building block for evolving Internet of Things (IoT) applications [36]. The main advantage of MCS is that data can be collected without much user intervention. In the present study, MCS focuses on quantifying the history of the exhibition of symptoms (such as decreased levels of physical activity, lesser social interactions, reduced mobility, etc.) over a period of time. We rely on the smart phone usage records collected for a specific duration of two weeks to analyse individuals' behavioural patterns.

This work presents experiments by integrating smart phone usage patterns with task-based experimental modalities. Smart phone usage patterns were utilized to present historical information of symptom exhibition. Task-based experiments called emotion elicitation (triggering emotions to activate visual manifestations by showing pictures/video clips) and speech elicitation (triggering auditory responses by reading a predefined paragraph/open form of speech) were conducted to observe and study the momentary representations.

The following are the main highlights/contributions of this work:

- To the best of our knowledge, this could be the first approach to integrate real-time smartphone usage patterns with the task-based modality for diagnosing depression.
- Creation of a novel tri-modality dataset with two weeks of smartphone usage data, visual and auditory cues of the participants.
- Designing an end-to-end machine learning pipeline, which involves multimodal data collection, feature extraction, feature selection, fusion, and classification to distinguish between depressed and non-depressed subjects.
- Extensive experimentation done using: various individual feature vectors from multi-modalities, features selection techniques, fused features, and machine learning classifiers such as Logistic Regression, Decision Tree(DT), Naive Bayes(NB), Random Forest(RF), Support Vector Machines (SVM), etc., for classification.
- Our findings demonstrate that the combination of statistical feature vectors from multimodal cues gave promising results compared to unimodal feature vectors, not only on our dataset but also on a benchmark open-source dataset.

The rest of the paper is organized as follows. Section II contains the details of related work, covering the existing works on various depression detection techniques. Section III presents the proposed approach. Section IV presents our findings and results in the results section. Section V presents the conclusion. Finally, Section VI presents limitations of the proposed approach and future works in the same field.

2 Related work

According to American Psychiatric Association, which has released the Diagnostic and Statistical Manual of Mental disorders-V (DSM-V), depression is a common mental disorder that involves a continuous sense of sorrow and/or distinct lack of interest. In addition to these, four or more following symptoms are present: weight loss or gain, sleep difficulties, i.e., insomnia or hypersomnia, psychomotor retardation, fatigue or loss of energy, diminished ability to think or concentrate, feelings of worthlessness or excessive guilt, and suicidal thoughts. Depression results in clinically notable changes in cognition, emotion regulation, or behavior that reflect the individual's psychological, biological, or developmental process, resulting in socially deviant behaviour. This condition persist for a minimum duration of two weeks [5].

Assessments are done through clinical consultations and questionnaire-based standard self-reports. Clinical consultations are conducted by psychiatrists, psychologists, experienced counsellors, etc. Table 1 gives details about a few standard self-reports. It usually takes 10 to 20 minutes to complete the questionnaire. Self-reports contain questions that are to be rated by an individual for the severity of their symptoms over a specific period of time. Each question records the response with 0 (not at all), 1(several days), 2(more than half the days), and 3(nearly every day). A score is formed by summing up all the responses. This score is used to diagnose depression and classifying the severity of the depression into different categories: mild, severe, etc.

Table 1 Few standard self-reports

Self-Report Questionnaire	No. of questions	Sample contents of the Questionnaires	Categories of depression
Patient Health Questionnaire(PHQ-9) [30]	9	Sleep difficulties, excessive guilt, fatigue, suicidal ideation	Mild, moderate, moderately severe, and severe
Beck Depression Inventory(BDI-II) [8]	21	Mood, self-hate, social withdrawal, fatigability	minimal, mild, moderate, and severe depression
Hamilton Rating Scale for Depression(HRS-D) [10]	17	Loss of interest, agitation, mood, loss of weight	Normal, mild, moderate, and severe depression
Quick Inventory of Depressive Symptomatology (QIDS) [51]	16	oncentration, suicidal ideation, sleep disturbance, self-criticism	Normal, mild, moderate, and severe depression

The following sub-sections briefly give an overview of the different works found in the field of depression diagnosis through facial, verbal, smart phone usage metadata, and multimodal cues. In every approach, the goals and various aspects such as the data collection process, data sources are different. Irrespective of these differences, each approach aims to explore innovative solutions that can assist in depression detection.

2.1 Depression detection through smart phone usage indicators

Some works on mobile crowd sensing have attempted to provide depression detection methods for the following reasons: First, most smartphones are equipped with multiple sensors that can continuously gather information about the users. This data can be monitored to understand behavioural patterns in real-time. Second, smartphones are unobtrusive, prevalent, and capable of data transmission to remote servers without requiring direct user interaction. Third, passive sensing applications that can run in the phone background to capture usage information and store it locally/server can be designed. Few are readily available, like SensusMobile [61], Funf journal application [1], etc. Fourth, smartphone-captured behavioral variations can be used as discriminative features for depression assessment. For e.g., people with depression are more likely to sleep lesser time than non-depressed people. This behavior pattern can be collected via brightness/light sensors present in the smartphone [41].

During a longitudinal study carried out by Masud et al. [37] in daily real-life scenarios, Inbuilt phone sensors such as the acceleration and Global Positioning System (GPS) sensor were used to classify physical activities and location movement patterns, respectively. Using a wrapper feature selection method, a subset of features were selected. Depression score was estimated using a linear regression model. SVM classifier was used to distinguish individual depression severity levels (absence, mild, extreme), with an accuracy of 87.2%.

Fukazawa et al. [20] collected raw sensor data from mobile phone, such as brightness, acceleration, rotation/orientation, and application usage. The author used them to form higher-level feature vectors. The fusions of these feature vectors were able to predict the

[1] https://www.funf.org/journal.html

stress levels among the participants. The results demonstrate that the combined features extracted from smartphone log data can be used to predict stress levels.

De Vos et al. [15] passively recorded geographical location data among healthy and depressed groups. From their results, it is evident that a strong correlation exists between geographical movements and depressed people.

2.2 Depression detection through facial indicators

Facial markers are extensively considered in depression diagnosis due to the following reasons: First, depressed individuals tend to have anomalous facial manifestations for e.g., fewer smiles, more frequent lip presses, prolonged activity on the corrugator muscle, sad/negative/neutral expression occurrence, fast/slow eye blinks, etc. Second, capturing visuals by web cameras has become effortless. Third, Several tools are now available to extract visual features, e.g., The Computer Expression Recognition Toolbox [35], OPENFACE [6], imotions[2] etc.

Wang et al. [58] have examined the facial cue changes between depressed and normal subjects in the same situation (while displaying positive, neutral, and negative pictures). To measure the facial cue changes on the face, they used person-specific active appearance model [11] to detect 68 point landmarks. Statistical features are extracted from distances between feature points of eyes, eyebrows, corners of the mouth to feed the SVM classifier. The classifier achieved 78% test accuracy.

Girard et al. [22] have studied the relationship between facial manifestations and how the severity of depression symptoms changes over time. During a clinical interview of a longitudinal study, they measured Action Unit's (AU) by Facial Action Coding System(FACS) [17, 18] between Low/High Symptom states. FACS has become the standard for muscle movements in the face. Each subtle muscle movement exhibited on the face is represented as AU. They found that AU 12 (Lip Corner Puller) is lower while AU 14 (Dimpler) is higher in a severe depressive state.

Alghowinem et al. [2] have observed that the eyelids' average distance (when opened) and duration of blinks vary between depressed and normal subjects. The findings conclude that depressed subjects tend to have a smaller average distance of the eyelids and duration of blink is higher than normal subjects. Alghowinem et al. [3] have also observed that head pose and movements significantly differ from depressed to normal people. They drew few conclusions: longer gaze time towards the right and down, slower head movements, and few head posture changes in depressed subjects.

2.3 Depression detection through verbal indicators

Acoustic features of speech play a vital role in diagnosis of depression for the following reasons: First, linguistic features (what subject speaks), paralinguistic features (how subject speaks), etc., are generally affected by the subject's mental state. Second, the clinician uses verbal indicators. Several studies have found distinguishable prosodic features such as pitch, loudness, energy, formants, jitter, shimmer, etc., between depressed and non-depressed individuals. Third, the ease of recording and availability of tools to extract the features such as openSMILE [19], PRAAT,[3] COVEREP [16], etc.

[2] https://imotions.com/

[3] https://www.fon.hum.uva.nl/praat/

Cummins et al. [12] have investigated good discriminative acoustic features that distinguish normal and depressed speakers. Features like Spectral centroid frequencies and amplitudes were computed using Mel-frequency Cepstral Coefficients (MFCC) then normalized. Multidimensional feature sets, i.e., combinations of those features have performed better when compared to single-dimensional features. They employed Gaussian mixture models to predict depressed and normal speakers. Further, Cummins et al. [13] analysed the effects of depression manifesting as a reduction in the spread of phonetic events in acoustic space. In their work, three acoustic variability measures: Average Weighted Variance (AWV), Acoustic Movement (AM), and Acoustic Volume, were used to model the trajectory of depressed speech in the acoustic space. They found that depressed groups often tend to have reduced vowel space when compared with healthy people.

Scherer et al. [54] have investigated reduced vowel space's association with the speech of individuals who exhibit depressive symptoms. They worked on a publicly available Distress Analysis Interview Corpus(DAIC) dataset [23]. They employed a voicing detection algorithm to detect voiced parts of the speech. COVERAP toolbox was utilized to track the first two formants (F1 and F2) in the voiced speech. Further, F1 and F2 were used to compute vowel space while uttering three kinds of vowel sounds i.e., /i/, /a/, and /u/. An unsupervised learning algorithm called K-means clustering(with k=12 and c=3) showed the association between vowel space and the depressed group.

2.4 Depression detection through multi-modal indicators

Recently some researchers have also tried to combine different modalities due to the following reasons: First, an individual modality's contribution can be better understood when the convergence of modalities is carried out. Second, each modality has its own advantages. Hence a combination can yield better outcome. Third, compatible characteristics of the features exist.

Williamson et al. [59] utilized feature sets derived from facial movements and acoustic verbal cues to detect psychomotor retardation. They employed Principal component analysis for dimensionality reduction and then applied the Gaussian mixture model to classify the combination of principal feature vectors.

Alghowinem et al. [4] showed that the fusion of different modalities gives an improvement when compared to the individual modalities at hand. Their aim was to develop a classification-oriented approach, where features were selected from head pose, eye gaze, and verbal indicators of the depressed and healthy groups. Classification of these feature sets achieved the best results through the use of the SVM classifier.

Williamson et al. [50] combined text, audio, and facial features to form hybrid fusion on a publicly available DAIC dataset [23]. Authors used deep learning for classification in thier study.

Most of the studies discussed in current section employ ML based methods (Support Vector Machines, Gaussian Mixture Models, Random Forest, etc.,) but not deep learning methods. For this insufficient training data availability could be the reason. ML based methods can be trained on lesser data, i.e., when compared with ML, deep learning needs larger training data [44]. Another reason could be supervised ML is more powerful when a known relationship exists between the inputs and labels. i.e., numerous features can be extracted and then evaluated to improve model accuracy.

3 Method and proposed approach

3.1 Overview

Figure 1 illustrates the overall architecture of the proposed approach. Our proposed approach has three stages; Stage 1: Data Collection, Stage 2: Feature Extraction, and Stage 3: Model Training and Testing.

First of all, multi modal data collection was done by performing a participatory mobile crowd sensing experiment (where real-time smart phone usage data was collected over a period of 2 weeks) and a task-based experiment (where 15 minutes of visual and auditory responses were recorded). Post data collection, standard self reports were used to collect the ground truth. The acquired multimodal data was then used in the feature extraction stage as follows - After data pre-processing, low-level features were extracted from multimodal data modalities(smart phone, audio-visual modalities). High-level features were formed using low-level features. The statistical feature vectors were extracted from the high-level features.The statistical feature vectors and ground truth labels were used in the model-building stage. In the model training, the feature vectors from individual and combination of data modalities were used as inputs for feature selection techniques in order to train ML classifiers such as Logistic Regression(LR), Decision Tree(DT), Naive Bayes(NB), Random Forest(RF), Support Vector Machine(SVM)). These classifiers were trained for classifying depressed and non-depressed classes of participants. The model training is done as one time process. For testing, the trained models were then used on the new test data to predict the participant's status.

3.2 Data collection

For mobile-sensor data collection, participants were volunteered through social networks, mailing lists, flyers, posters, and personal contacts. Among 143 responses received, 102 participants (56% female, mean age of 18-19 years) met the experiment's eligibility criteria. Participants were eligible if they had smartphones, with access to the internet, could speak and read English, were over 18 years old, and lived in India.

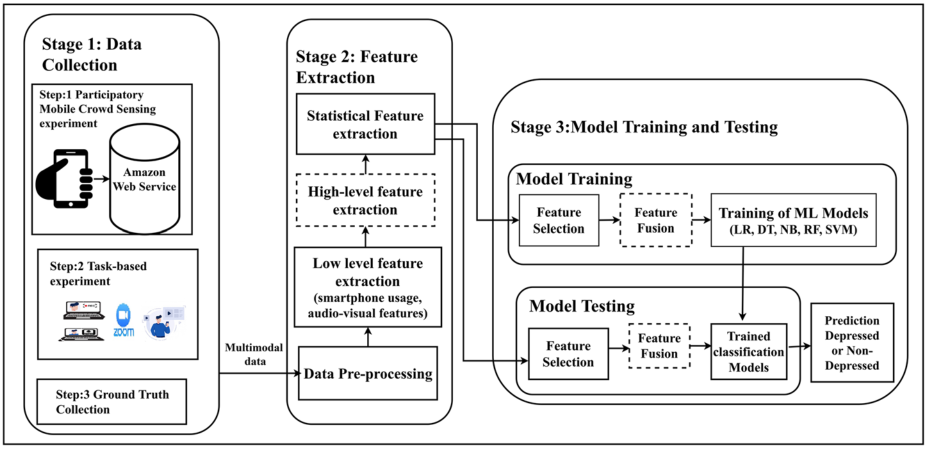

Stage 1: Data Collection
Step:1 Participatory Mobile Crowd Sensing experiment
Amazon Web Service
Step:2 Task-based experiment
zoom
Step:3 Ground Truth Collection
Multimodal data
Stage 2: Feature Extraction
Statistical Feature extraction
High-level feature extraction
Low level feature extraction (smartphone usage, audio-visual features)
Data Pre-processing
Stage 3:Model Training and Testing
Model Training
Feature Selection
Feature Fusion
Training of ML Models (LR, DT, NB, RF, SVM)
Model Testing
Feature Selection
Feature Fusion
Trained classification Models
Prediction Depressed or Non-Depressed

Fig. 1 The overall architecture of the proposed approach

To improve data quality, several incentives were provided to the participants to participate with seriousness. The data was collected for over a period of two weeks.

Two days before the start of the study, The research team assisted with the download, installation, and configuration of the SensusMobile application (refer to the About Sensus-Mobile sub-section below). All the participants were instructed to: keep their phones with them charged throughout the day and enable the GPS and Bluetooth sensors for the duration of the study. Before taking informed written consent from the participants, the purpose of data collection and explanation of data was provided.The research team periodically checked the data at the server and contacted the participants in case of any discrepancies.

At the end of the mobile-data collection phase, task-based data collection was performed in online mode to adhere Covid 19 guidelines and to accommodate participants across various regions. Participants were given appointments for Face to Face (30 minutes) Zoom sessions with members of the research team. The participants were instructed to turn their camera ON and be present in optimal lighting conditions with their headset before joining the session, preferably on a Desktop and if not on mobile. During the session, the researchers conducted tasks involving emotion [57] and speech elicitation [7] to record each subject's facial and acoustic responses. Table 2 lists the experimental procedure with time duration involved to conduct emotion and speech elicitation.

During online task-based experimental data collection, research assistants shared their screen/audio to perform emotion elicitation. In this task, various kinds (positive/neutral/sad) of multimedia clips selected from famous film clips in psychology [57] were shown to evoke the participant's emotions. Prior to each clip, the experimenter stated that the screen would be blank for one minute (when participants were asked to clear their minds of all feelings, memories, and thoughts). After all the clips were presented, the participants were provided a break of approximately 10 minutes, then speech elicitation was performed. The participants were asked to provide their speech in two different conditions. In the first scenario, they were asked to read out tale (a phonetically balanced paragraph called "The North and the South Wind") from the screen. Secondly, they were asked to provide an impromptu report

Table 2 Experimental procedure with time duration to conduct emotion and speech elicitation

Experimental tasks	Procedure	Description(source)	Duration
	Blank screen	NA	1 minute
	Positive video	The Circus(1928) / Charlie Chaplin, a known comedian, performs hilarious acts when he enters a lion cage.	3:32 minutes
Emotion Elicitation	Blank screen	NA	1 minute
	Neutral video	Abstract Shapes/colour bars	3 minutes
	Blank screen	NA	1 minute
	Negative video	The Champ(1979) / Little boy crying when his father is on the death bed.	3 minutes
Break	Blank screen	NA	10 minutes
Speech Elicitation	Passage reading	Short tale called "The North and the South Wind"	1 minute
	Free form speech	Participant's choice from a list appears on the monitor	2 minutes

on a topic of their choice from a list that appeared on the screen (e.g., memorable incident in life, their goals, etc.). The session was recorded for data pre-processing.

3.2.1 Ground truth labelling

The SensusMobile app was programmed to deliver instances of the PHQ-4(a subset of PHQ-9) survey (to be filled by participants) on a daily basis and a Patient Health Questionnaire(PHQ-9) Online Survey at the end of the mobile-data collection period (two weeks). PHQ-9/4 were selected in our study due to their high levels of consistency and statistical reliability/validity. Kroenke et al. [30] conducted a study to examine the validity of PHQ-9 for depression assessment. Their findings suggest that PHQ-9 is reliable/valid, and it is a helpful research tool for depression diagnosis. The participants also provided physical copies of completed PHQ-9 responses during the video-data collection phase. This questionnaire was collected in order to obtain a more “current” representation of the participants’ psychological state (compared to the mobile data collection phase). Based on the PHQ-9 Scores, each participant’s mental status was categorized into binary labels (Non-depressed=0; those who show depressive symptoms =1).

To take care of outliers or inappropriate filling on the PHQ-9 report by the participants, a team of psychology research scholars from the Central University of Tamil Nadu were provided with scanned copies of the PHQ-9 questionnaires and the recordings of the participants(attained during task-based experimental data collection). The scholars reviewed these items and provided binary classifications using an amalgamation of both resources provided as well as another interview with the participant if needed. This additional step verification was performed in order to strengthen the validity of the ground truth labels leading to a more coherent dataset. Overall data set has been labelled with 54 non-depressed and 48 depressed subjects.

3.2.2 About sensusmobile

SensusMobile [61] is an open-source mobile crowd data collection application that runs in the background to access readings from device hardware sensors (i.e., accelerometer, GPS, gyroscope, etc.). The accessible sensor information can be stored locally on the device or transmitted to a remote server. Our project utilized Amazon Simple Storage Service Web Services for data collection to ensure that participant data gets stored periodically without user intervention. Prior to the actual study, several beta tests were performed with SensusMobile to calibrate settings to minimize battery consumption and reduce data redundancy.

SensusMobile supports two methods of data sensing: 1) Listening (continuous data collection) and 2) polling (periodical triggering of probes to collect readings). Table 3 lists a subset of the data items collected from the participant’s smartphone for the study conducted.

3.3 Data pre-processing

The goal of the data pre-processing stage was to facilitate the extraction of features from various sources, i.e., Smartphone usage data(.json) files along with visual and speech recordings.

The research team extracted visual cues (.mp4) from participants via recorded visuals as marked in Fig. 2(a), and the marked red box was cropped for further use. Video portions where the participant is not visible or people other than the participant appeared in

Table 3 A subset of the data items collected from the participant's smartphone for the study conducted

Data Collected	Probe Used	Listening/polling	Intervals
Acceleration	Accelerometer	Listening	1 reading / second
Application Usage Statistics	ApplicationUsageStats	Polling	1 reading / 15 min
Brightness	LightDatum	Listening	1 reading / second
Bluetooth encounters	BluetoothDeviceProximityDatum	Polling	scans and reads performed for 10 seconds each, between 30-second intervals
Gyroscope values	GyroscopeDatum	Listening	1 reading / second
GPS/ location	LocationDatum	Polling	1 reading / 15 min
Screen unlocks	ScreenDatum	Polling	1 reading / 30 seconds

the recording were manually deleted. Openface [6] was used to extract low-level features from cropped versions of the recording. It is a state-of-the-art framework for extraction of low-level features like facial landmark location detection, eye gaze estimation, head pose estimation, and facial action unit recognition. In the proposed approach, higher-level statistical feature vectors were formed from these low-level features.

From the recorded meeting of the speech elicitation phase, as shown in Fig. 2(b) speech cues (.mp3) of the participants were extracted. The SOX tool[4] was used for noise removal from the extracted speech content. Then, the Praat software tool[5] was used for low-level acoustic feature extraction (pitch, intensity, formants, etc.).

3.4 Feature extraction

Those features extracted in the proposed approach were considered clinically significant [5] and supported by related work. It is believed that features computed from entire raw data obtained during the data collection of each modality give insightful information than on samples of information. For example, smart phone usage feature extraction with 14 days of smart phone usage data is more insightful than 3/7/10 days of data. In the following sub-sections, feature extraction is explained in detail.

3.4.1 Smart phone usage feature extraction

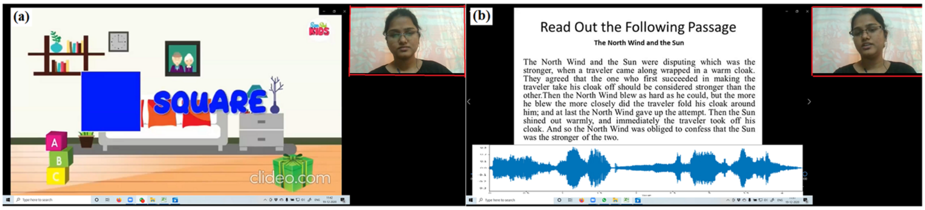
The following features were computed from collected smart phone usage data. Although clinically, those features extracted in the proposed work have no direct relation to depression, they can help quantify the individuals' physical, cognitive, and environmental levels [55]. Features are as follows:

Accelerometer probe features

The accelerometer records dynamic or static forces the sensor is experiencing in x, y, and z directions. The acceleration probe reading consists of x, y, and z axes, which specify the

[4] http://sox.sourceforge.net/

[5] https://www.fon.hum.uva.nl/praat/



Read Out the Following Passage

The North Wind and the Sun

The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveler fold his cloak around him; and at last the North Wind gave up the attempt. Then the Sun shined out warmly, and immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

Fig. 2 Samples during elicitation methods (a) Emotion Elicitation: participant's facial clues are recorded while watching the neutral video (b) Speech Elicitation: participant's speech is recorded while reading the phonetically balanced paragraph

axes' acceleration. In this study, tri-axis readings were considered for feature extraction. From the raw three-axis accelerometer data, which was taken at 1800 samples per hour, the accelerometer magnitude was computed using (1). Further arithmetic mean of accelerometer magnitude was also computed from the accelerometer readings [27, 28].

$$\text{Magnitude} = \sqrt{x_i^2 + y_i^2 + z_i^2} \tag{1}$$

Where x_i, y_i, z_i are the accelerometer readings at a given time instant i.

Gyroscope probe features

The Gyroscope sensor aids in determining the orientation of a device using the earth's gravity. It tracks the rotations of the device in x,y and z directions. The gyroscope probe reading consists of x, y, and z axes, which specify the axe's rotations. From the raw individual axis gyroscope data taken at 1800 samples per hour, the Variance of the individual axis was computed.

Application usage probe features

Smartphone applications were clustered into various categories from their google play store website entries. For Example- WhatsApp is part of the "communication" category

The average amount of hours spent on each category per day was computed (using the TimeInForeground entry (provided by the sensor) as a variable).

The subset of categories considered for the study were communication category, social category, entertainment category, health and fitness category, music and audio category, weather category, travel category, books and reference category, shopping category, events category, photography category, maps and navigation category, business category, etc.

Location probe features

Location probe, also known as GPS, periodically records the latitude and longitude entries of the user. Four samples per hour were collected. Using these samples, location variance [52] was calculated as the combined Variance of the latitude and longitude components as shown in (2).

$$\text{Location Variance} = \log(\sigma_{lat}^2 + \sigma_{long}^2) \tag{2}$$

Where σ_{lat}^2, σ_{long}^2 are the Variance's of latitude and longitude, respectively. Speed mean [52] was also extracted, i.e., mean of instantaneous speed's obtained at each location sample.

The instant speed was computed as the change in latitude and longitude values over two consecutive instants, as shown in (3)

$$\text{Speed mean} = \sqrt{(\frac{lat_i - lat_{i-1}}{t_i - t_{i-1}})^2 + (\frac{long_i - long_{i-1}}{t_i - t_{i-1}})^2} \tag{3}$$

Where $lat_i, long_i$ are the latitude and longitude at the time of sample i.

Variance and mean were computed on instantaneous speed values along with total distance [52] -i.e. total geographical displacement using (4)

$$\text{Total distance} = \sum_i \sqrt{(lat_i - lat_{i-1})^2 + (long_i - long_{i-1})^2} \tag{4}$$

Where $lat_i, long_i$ are the latitude and longitude at the time of sample i.

Bluetooth probe features

Bluetooth probe was used to track nearby devices using Bluetooth via periodic information transmission. There were 12 samples recorded per hour, and the probe components were grouped day wise and the number of unique encounters was computed using addresses as an index. Features such as Average Unique Bluetooth encounters per day, Variance, and Standard Deviation of the number of unique devices encountered were used in the study.

Light/brightness probe features

The Light probe measures the illuminance of the device (user). The app recorded light levels 12 times per hour. Mean, Variance, and Standard Deviation of the readings were chosen as features.

Screen unlock probe features

The Screen Unlock probe is activated whenever the user unlocks his/her screen. This probe returns true (Boolean value) once the user unlocks the screen and the time stamp is recorded. The total count of all screen unlocks has been used as a feature [9, 40].

Table 4 summarizes all the smart phone usage features.

3.4.2 Visual feature extraction

Two kinds of Openface's low-level features were used for feature extraction. They are 1) 68 facial landmark location coordinates, their visualization is shown in Fig. 3(a). 2) Facial Action Coding System(FACS). FACS is a method for categorizing facial movements based on their appearance on the face. The facial Action Unit (AU) represents almost every subtle movement of muscles on the face. Figure 3(b) shows a few AU's. A subset of AU (specifically: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, and 45) are recognized by Openface. Each AU has the following - AU occurrence (0 if the AU is absent and 1 if it is present) and AU intensities (degree of variability in the scale of 0 to 5, where 0,1 and 5 represent not present, minimum and maximum intensity, respectively).

In the present study, visual features are of two categories: 1) Geometrical features(using 68- landmark locations) and 2) Facial Action unit features(using FACS). Different geometrical features, i.e., displacement features, distance features, and region unit features were formed. Further statistical features were extracted from these two categories of features.

Table 4 Smart phone usage features

Parent feature	Description	Statistical features extracted	No. of features
Accelerometer probe	To measure acceleration (the rate of change of velocity). We approximated accelerometer magnitudes using (1).	Mean of the accelerometer magnitude was computed	1
Gyroscope probe	To measure orientation of the phone.	Axis-wise variance of entries were calculated	3
Application Usage probe	App categories were extracted using their package references from play store	Average amount of hours per day spent on each application category by the user	36
Location probe	Raw readings were used to calculate location variance (2), speed mean (3) total distance (4).	location variance (2) and its mean, Variance and mean of the instantaneous speed (3) and total distance (4) were calculated.	6
Bluetooth probe	The entries are grouped day wise and the number of unique encounters were calculated using the Address entry.	Day-wise mean, Variance and Standard Deviation of entries	3
Light (brightness) probe	To measure the illumination of the device(user). Brightness probe readings are used here	Mean, Variance and Standard Deviation of readings	3
Screen unlock probe	The entries were divided on the basis of binary readings provided by the probe	The percentage of entries where the screen_on entry is True with respect to the total number of entries was calculated	1

Geometrical features

Displacement Features

Using one coordinate, the displacements of 6 specific landmark points (marked as dp1,dp2,dp3, dp4,dp5 and dp6 in Fig. 4) were computed using (5)

$$\text{Displacement} = \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2} \tag{5}$$

where (x_i, y_i) denotes the landmark coordinates present in the frame i, (x_{i+1}, y_{i+1}) the same landmark coordinates in the frame $i + 1$ where i ranges from 0 to $n - 1$ (0 is the first and $n - 1$ is the last frame).

Distance features

Using two coordinates, Eight Euclidean distance values i.e mean squared distances (marked as d0,d1,d2,d3,d4,d5,d6, and d7 in Fig. 4) between two pairs of coordinates were computed using (6).

$$\text{Distance} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{6}$$

where (x_i, y_i) and (x_j, y_j) represents landmarks of two different coordinates in the same frame. These distances were calculated for all the frames.

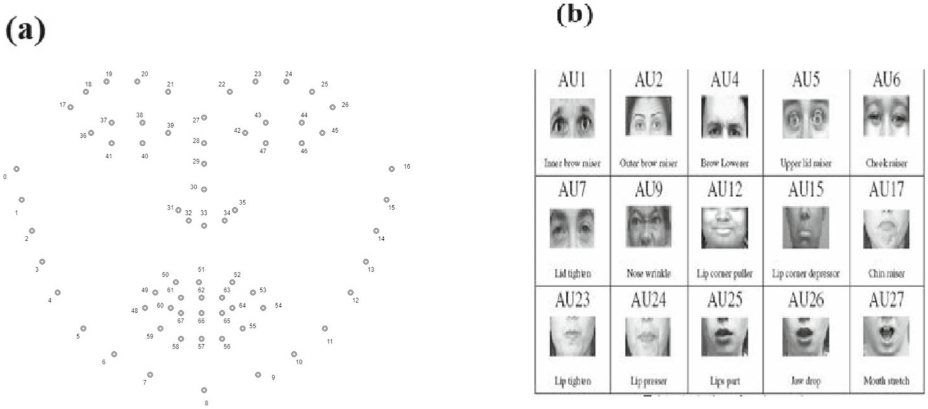


Fig. 3 Visual feature extraction (a) Visualization of 68 facial landmark location coordinates (b) Examples of few action units extracted from Cohn and Kanades database [25]

Region unit features

Using more than two coordinates(as marked as A0, A1, and A2 in Fig. 4), the Area of the irregular polygon was used to compute the area of the mouth, left eye, and right eye using specific points in that region using (7)

$$\text{Area} = \frac{1}{2}\left|\sum(x_i y_{i+1} - x_{i+1} y_i)\right| \tag{7}$$

when $i = n - 1$, then $i + 1$ is expressed as 0. where (x_i, y_i), (x_{i+1}, y_{i+1}) to (x_{n-1}, y_{n-1}) represents the set of points in frame i.

Facial action unit features

Features from each AU occurrence and AU intensities were taken.

From both: geometrical and action unit features, statistical features like mean, median, standard deviation, etc., listed in Table 5 were extracted.

3.4.3 Audio feature extraction

Generally, depression diagnosis is subjective in nature which can be manipulated. So we assumed that acoustic features are more powerful than linguistic characteristics. Audio features were computed from the audio files which were recorded at the sampling frequency of 32000 Hz during the speech elicitation experiment. Table 6 lists the details of the features that were extracted.

3.5 Feature selection

Feature Selection is a mechanism to choose an optimum subset of features that improves classification efficiency with less complexity and computing costs. In the current study,

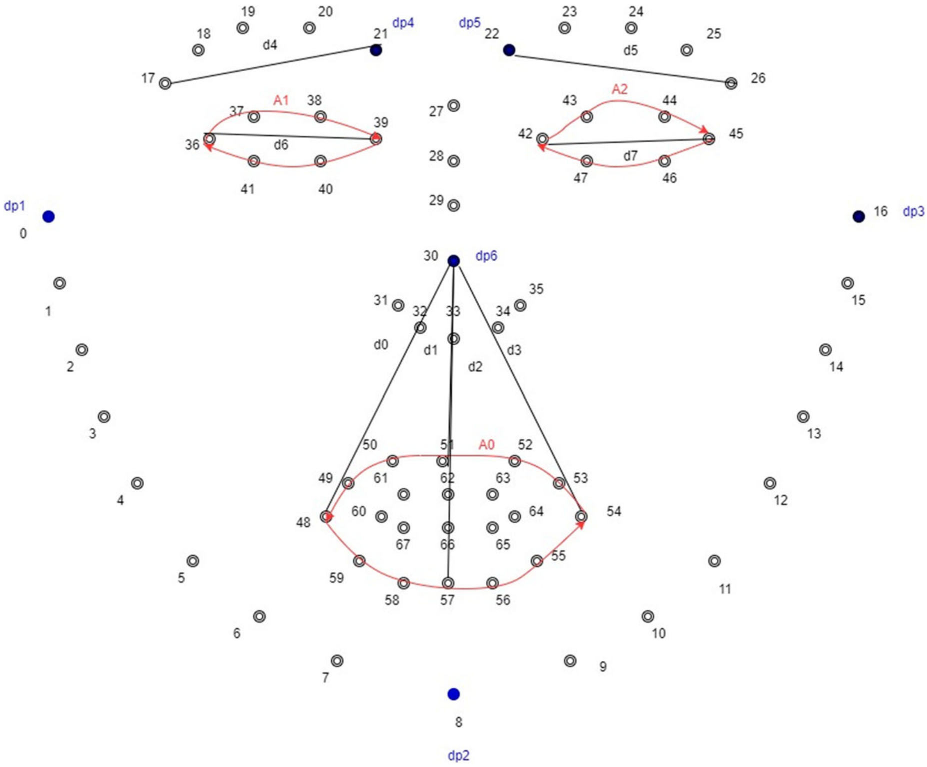


Fig. 4 Geometrical Features representation using facial landmark locations

numerous features (smart phone usage-53,visual-413, and audio-82) were extracted. Generally, high dimensionality of input features may lead to poor performance because feature space becomes huge and also it is observed that our dataset contains correlated features. Therefore, two different feature selection approaches were experimented to select the better features which improves the accuracy.

3.5.1 Feature selection using correlation

Correlation analysis is a statistical technique used for measuring the strength of the linear relationship between two or more attributes [1]. The Pearson correlation coefficient technique was used in the present study for three reasons:1) it is easy to implement,2) it suits our data (type), and 3) it is vastly used in the literature for depression detection [31, 33, 38, 39, 49, 60].

Given two attributes X and Y, which have ‘n’ values. The Pearson’s correlation coefficient (r) can be determined using (8).

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{8}$$

Table 5 Summary of the facial features

Feature	Feature	Description	Statistical features	No. of
category	Name		extracted	features
	Displacement features	Displacement (using (5)) of the six specific points as marked in Fig. 4 denoted by blue points as dp1 to dp6.	Mean, median, minimum, maximum, kurtosis, mode, standard deviation, Root mean square, skewness for Each of 6 displacement points. (dp1 to dp6)	54
Geometrical features	Distances features	Distances(using (6)) between 8 pairs of points as marked in Fig. 4 denoted by black lines as d0 to d7	Mean, median, minimum, maximum, kurtosis, mode, standard deviation, Root mean square, skewness for Each of 8 distances. (d0 to d7)	72
	Region Units	Area of the mouth. Area of the left eye and Areaof the right eye(7) as marked in Fig. 4 is denoted by red irregular lines as A0, A1 and A3.	Mean, median, minimum, maximum, kurtosis, mode, standard deviation, Root mean square, skewness for Areas of mouth, left eye and right eye. (A0 A1 and A2)	27
Facial Action Unit Features	Action Unit features	The facial action coding system is used to quantify the muscle movements on the face. AU occurrences present(1) or absent(0) for 18 AU.	Mean, median, standard deviation, kurtosis, mode, Root mean square, skewness for each 18 AU present/ absent.	126
		If present, AU intensities for 17 AU intensities.	Mean, median, standard deviation, maximum, kurtosis, mode, Root mean square, skewness for each 17 AU intensities	136

Table 6 Audio features

Feature Name	Description	Statistical features extracted	No. of features
Pitch	It is an approximation of the quasi-periodic rate of vibrations per speech cycle.	mean, median, standard deviation, minimum, mode maximum, kurtosis, Root mean square, skewness	9
Intensity	It is the measure of the perceived loudness.	mean, median, standard deviation, minimum, mode maximum, kurtosis, Root mean square, skewness	9
Formants [F1,F2, F3,F4]	They indicate resonating frequencies of the vocal tract. The formant with the lowest frequency band is F1, then the second F2, which occurs with 1000Hz intervals.	mean, median, standard deviation, minimum, maximum, kurtosis, mode, Root mean square, skewness	36
Pulses	A fundamental, audible, and steady beat in the voice.	Count, Mean, standard deviation, variance	4
Amplitude	It is the size of the oscillations of the vocal folds due to vibrations caused by speech biosignal.	minimum, maximum, mean, Root mean square	4
Mean Absolute jitter	It is the absolute difference between consecutive vocal periods, divided by the mean vocal period.	Mean	1
Jitter (local, absolute)	The absolute difference between consecutive periods, in seconds.	Mean	1
Relative average perturbation jitter	It measures the effects of long-term pitch changes like slow rise/fall in pitch. It is calculated as the average absolute difference between a period and its average and its 2 neighbours, divided by the mean period.	Mean	1
5-point period perturbation Jitter	It is calculated using the average absolute difference between a period and the average of it and its 4 closest neighbours, divided by the mean period.	Mean	1
Mean absolute differences Jitter	It is the absolute difference between consecutive differences between consecutive periods, divided by the mean period	Mean	1

Table 6 (continued)

Feature Name	Description	Statistical features extracted	No. of features
Shimmer	It defines the short-term (cycle-to-cycle) tiny fluctuations in the amplitude of the waveform which reflects inherent resistance/noise in the voice biosignal.	Mean	1
Mean Shimmer	Average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.	Mean	1
Mean Shimmer dB	average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20.	Mean	1
3-point Amplitude Perturbation Quotient Shimmer	It is calculated as the average absolute difference between the amplitude of a vocal period and the average of the amplitudes of its neighbours, divided by the average amplitude.	Mean	1
5-point Amplitude Perturbation Quotient Shimmer	It is the average absolute difference between the amplitude of a vocal period and the average of the amplitudes of it and its 4 closest neighbours, divided by the average amplitude.	Mean	1
11-point Amplitude Perturbation Quotient Shimmer	It is the average absolute difference between the amplitude of a vocal period and the average of the amplitudes of it and its 10 closest neighbours, divided by the average amplitude	Mean	1
Mean absolute differences shimmer	Average absolute difference between consecutive differences between the amplitudes of consecutive periods.	Mean	1
Harmonicity of the voiced parts only	It is used for measuring the repeating patterns in voiced speech signals.	Mean	1
Mean autocorrelation	It is used for measuring the repeating patterns in the speech signal.	Mean	1

Table 6 (continued)

Feature Name	Description	Statistical features extracted	No. of features
Mean harmonics-to-noise ratio	It is a measure which gives the relationship between the periodic and additive noise components of the speech signal.	Mean	1
Mean noise-to-harmonics ratio	It is a measure which gives the relationship between the periodic and additive noise components of the speech signal.	Mean	1
Fraction of locally unvoiced frames	It is a fraction of pitch frames analysed as unvoiced pitch (75Hz) frames in a speech biosignal of a specified length.	Mean	1
Number of voice breaks	The number of distances between consecutive vocal pulses that are longer than 1.25 divided by the pitch floor. Hence, if the pitch floor is 75 Hz, all inter-pulse intervals which are longer than 16.6667 ms are called as voice breaks.	Count	1
Degree of voice breaks	This measure is the total duration of breaks between the voiced parts of the speech signal.	Mean	1
Total energy	Total energy of a vocal signal in air.	Mean	1
Mean power	The mean power of a speech signal in air.	Mean	1

r is the correlation coefficient, $Cov(X, Y)$ is the covariance, $\sigma_X$and σ_Y are the standard deviation of X and Y, respectively. Suppose X and Y are two set of values containing $[x_1, x_2, \cdots x_n]$ and $[y_1, y_2, \cdots y_n]$. r value can be calculated using (9)

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{9}$$

where n is the size of the sample, X_i and Y_i are the i^{th} data value and $\bar{X}$, $\bar{Y}$ are the mean values of X, Y respectively.

r value lies in between -1.0 and +1.0. The r defines two parameters for any given two sets of values. They are:1) strength: it measures how much these sets are associated, higher the value greater the relationship) and 2) direction of the relationship: if one value increases in one set, then other value increases in another set or one value decreases in one set, then other value decreases in another set. In short, when both move in the same direction. it is called positive correlation. Converse is a negative correlation, i.e., in the opposite direction. 0 signifies there is no relationship. As the value reaches closer to +1, the relationship becomes stronger.+1 indicates a perfect strong correlation. Similarly, those values close to -1 show a strong negative correlation. Figure 5 shows visualization of three kinds of correlations.

In the present work, the r values between the features and labels were examined. While examining, we observed a high correlation in the features themselves of our dataset. So, we decided to reduce the redundancy of the dataset by removing the highly correlated features. In all such cases, only one feature with a high correlation value with the label was selected, while other features were removed from the dataset. After deriving feature sets into the training and testing, a threshold correlation value(r) of 85% in the training data set was selected by experimentation(which resulted in improvement in terms of accuracy), and then resultant features were dropped in both the training and also testing dataset (to avoid overfitting).

3.5.2 Feature transformation using dimensionality reduction

A feature transformation technique called Principal Component Analysis (PCA) was used. PCA was applied to reduce/minimize the dimensions of the feature vector. The number

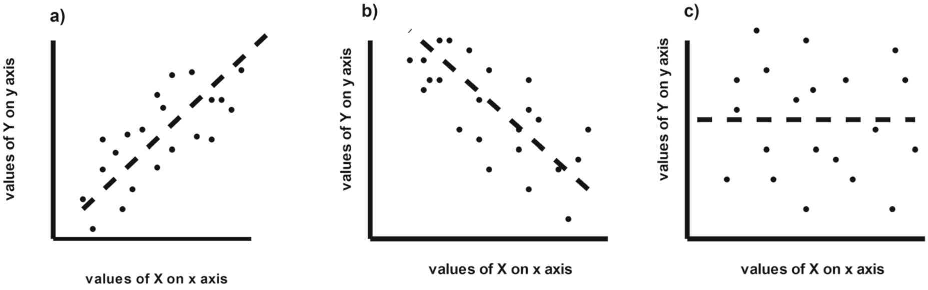


Fig. 5 Visualization of three kinds of correlation :a) positive correlation b) negative correlation, and c) no correlation

of components in the resultant feature vector was based on the most promising principal components that have 95% variances to classify the labels.

3.6 Normalization

Each modality in our study belongs to various scales. Hence min-max normalization, i.e., scaling between 0 and 1 on the feature vectors on individual modality, was applied. The current study was performed on the normalized feature vectors. Figure 12(a) for visual representation.

4 Results

In this section, first, the efficacy of the extracted features using statistical analysis of Pearson's correlation coefficients is described. Second, the classification results using ML classifiers (LR, DT, NB, RF, and SVM) on individual data modality and fused data modalities are presented. Lastly, the effectiveness of the proposed approach is demonstrated by comparing the results on a subset of feature vectors of a benchmarking dataset in depression detection called the Distress Analysis Interview Corpus (DAIC) dataset [23].

4.1 Statistical analysis

This subsection describes the efficacy of the extracted features using statistical analysis to prove the capability of the features to predict the depressed or non-depressed subjects. Each feature value and corresponding binary class label was analysed using pair-wise comparison with Pearson's correlation analysis. The r values(using (9)) were computed using this pair-wise comparison to find positive and negative correlated features. These r values were sorted to pick the top 10 correlated features in each category.

Table 7 lists the top 10 features which were found to be positively correlated with the ground truth labels. It is worth noting that all the top 10 features are Action Unit (AU) features of the visual modality. Table has the AU feature name, its description, and r value (the

Table 7 The top 10 positive correlated features with their description, r value (strength of correlation and direction)

S.no	Feature Name	Description	r value
1	AU 12 standard deviation (A_12_S)	Lip corner puller intensity Standard deviation	0.64277
2	AU 12 root mean square (A_12_R)	Lip corner puller intensity root mean square	0.62708
3	AU 12 maximum (A_12_M)	Lip corner puller intensity maximum	0.562378
4	AU 12 mean (A_12_MN)	Lip corner puller intensity mean	0.51846
5	AU 10 standard deviation (A_10_S)	Upper lip raiser standard deviation	0.51244
6	AU 06 maximum (A_6_M)	Cheek Raiser maximum	0.49279
7	AU 25 root mean square(A_25_R)	Lips part root mean square	0.48731
8	AU 25 count (A_25_C)	Lips part count	0.48315
9	AU 25 mean (A_25_M)	Lips part mean	0.47884
10	AU 06 standard deviation (A_6_S)	Cheek raiser standard deviation	0.473363

strength of the correlation and direction). Table 8 lists the top 10 features which are negatively correlated. it contains feature name, its description, and r value. We have listed only 10 features for simplicity because the features extracted were numerous in the conducted study.

Figures 6 and 7 show the positive and negative correlations in the sample of participants, respectively. The feature vectors are normalized (0 to 1) for better understanding. To avoid clutter in the graphs, we have chosen only top five features rather then all the top 10 features listed in both categories.

The graph in Fig. 6 shows the variations of positively correlated features between depressed and non-depressed subjects. For example, A_12_R (see Table 7 S.No 2) has lower values in depressed (1-5) and higher values in Non-depressed (6-10).

The graph in Fig. 7 shows the variations of negatively correlated features between depressed and non-depressed subjects. For example, F_L_U(see Table 8 S.No 2) has higher values in depressed (1-5) and lower values in Non-depressed (6-10).

Figures 8 and 9 show the participant's wise variations in the positive and negative correlated features for a sample of participants, respectively.

The graph in Fig. 8 shows insights into how the positively correlated features vary between depressed and non-depressed subjects. For example, Depressed subjects exhibit lower values in A_12_S (see Table 7 S.No 1) features and higher values for non-depressed subjects for the same feature.

The graph in Fig. 9 shows insights into how the negatively correlated features vary between depressed and non-depressed subjects. Depressed subjects exhibit higher values in A_25_S (see Table 8 S.No 1) feature and lower values for non-depressed subjects for the same feature.

Table 8 The top 10 negative correlated features with their description, r value(strength of correlation and direction)

S no	Feature Name	Description	*r* value
1	AU 25 skewness (A_25_S)	Lips part count skewness	-0.44741
2	Fraction of locally unvoiced frames (F_L_U)	It is a fraction of pitch frames analyzed as unvoiced pitch (pitch is 75Hz) frames in a voice.	-0.3891
3	Degree of voice (D_V_B)	This measure is total duration of breaks between the voiced parts of the speech signal	-0.3784
4	AU 10 skewness (A_10_SK)	Upper lip raiser skewness	-0.36990
5	AU 09 skewness (A_9_S)	Nose wrinkle skewness	-0.34867
6	AU 25 kurtosis (A_25_K)	Lips part kurtosis	-0.34069
7	Pitch Skewness (P_SK)	It is pitch's skewness	-0.3217
8	AU 12 skewness (A_12_SK)	Lip corner puller skewness	-0.3195
9	Shimmer APQ3 (SAQ)	It is the average absolute difference between the amplitude of a vocal period and the average of the amplitudes of it and its 2 closest neighbours, divided by the average amplitude.	-0.312
10	Mean absolute differences shimmer(MDS)	Average absolute difference between consecutive differences between the amplitudes of consecutive periods.	-0.3129

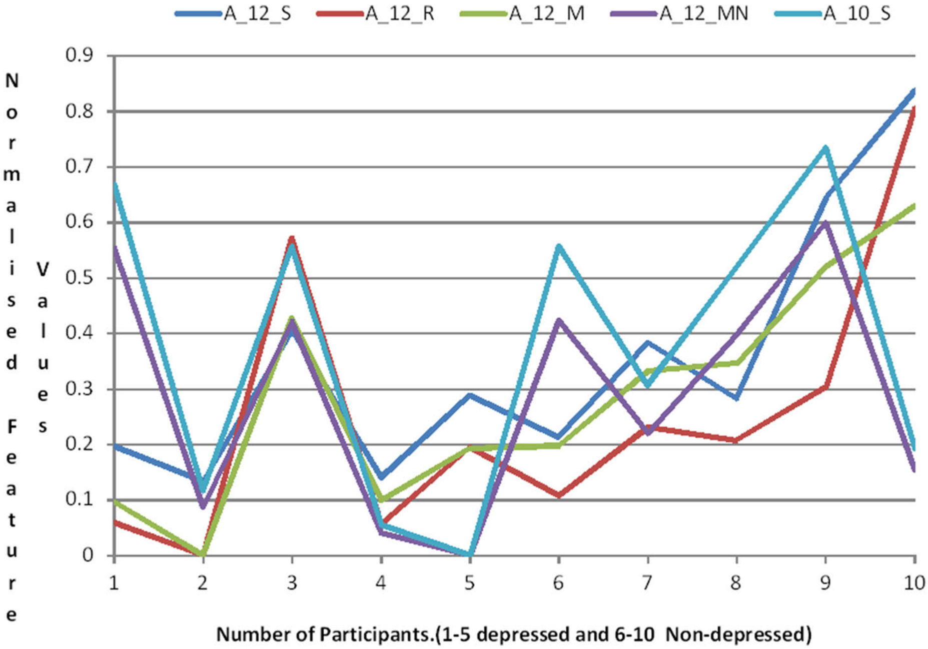


Fig. 6 Top 5 Positive correlated feature variations

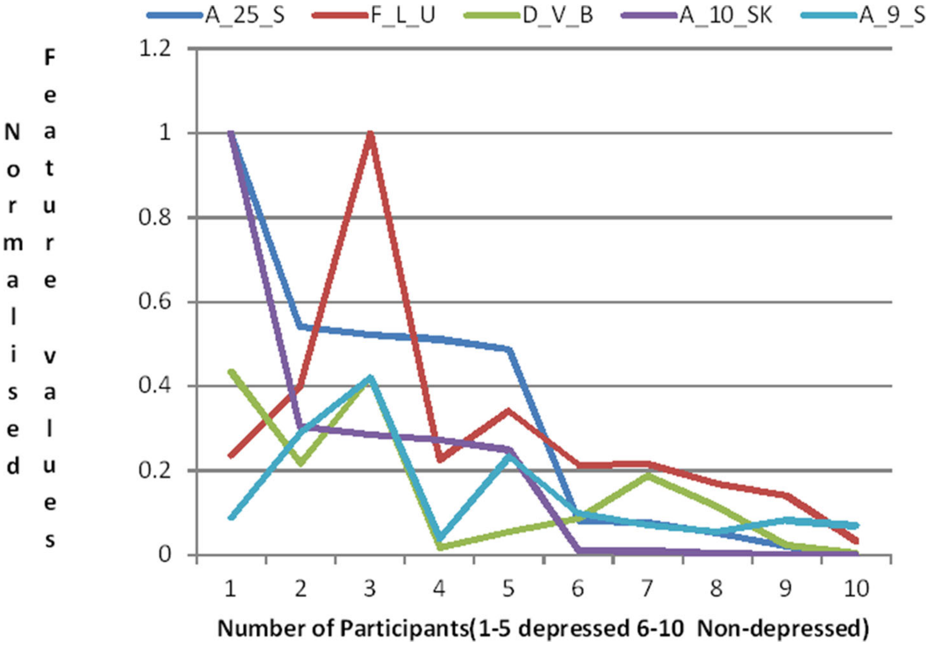


Fig. 7 Top 5 Negative correlated feature variations

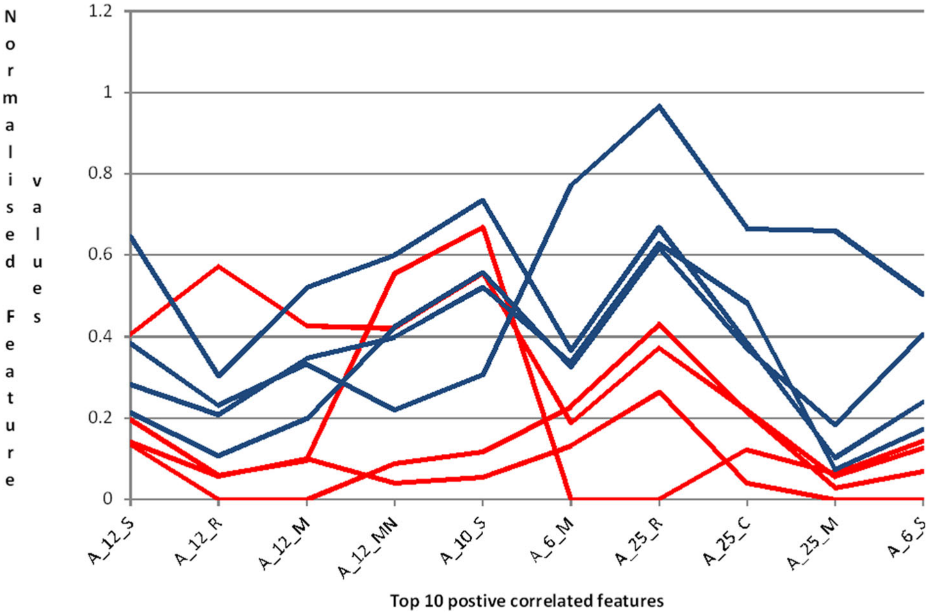


Fig. 8 Participant wise variations in Top 10 positive correlated features. Red and blue lines indicate the depressed and non-depressed participants, respectively

To show single feature variations in all the participants, A_10_S (see Table 7. S.No 5) feature from the positive correlated feature set and A_25_S (see Table 8 S.No 1) feature from the negative correlated feature set were selected. Figures 10 and 11 show single feature variations in positive and negative correlated features, respectively.

The graph in Fig. 10 shows how values are different in a single feature(positive correlated) between non-depressed and depressed subjects. For example the values of A_10_S (see Table 7. S.No 5) feature have higher values for the most non-depressed subjects and lower for the depressed subjects.

The graph in Fig. 11 shows how values are different in a single feature(negative correlated) between non-depressed and depressed subjects. For example, the values of A_25_S (see Table 8 S.No 1) feature have lower values for the most non-depressed subjects and higher values for the depressed subjects.

4.2 Classification results of individual modality and using feature fusion

This subsection describes the classification results using a family of machine learning classifiers implemented on individual data modalities and by fused data modalities. ML classifiers like LR, DT, NB, RF and SVM were used with default hyper parameters. All the results presented are in terms of average accuracy because of the balanced dataset. The dataset was randomly categorised (without any overlap) into two components:80 percent training data and 20 percent testing data. Table 9 lists the classification results using individual modality (see Fig. 12b for visual representation) with:1) All the extracted feature vectors (see row# 1 to 5) then with feature selection mechanisms using the reduced feature vectors based on 2) Pearson's statistical correlation analysis(see row# 6 to 10), and 3) PCA (see row# 11 to 15).

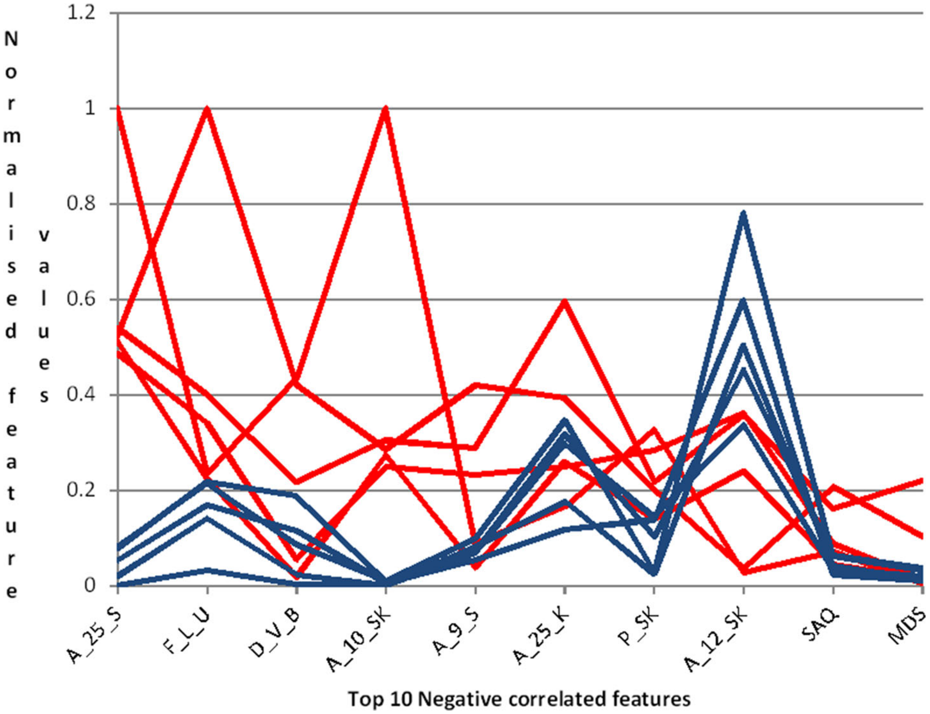


Fig. 9 Participant wise variations in Top 10 negative correlated features. Red and blue lines indicate the depressed and non-depressed participants, respectively

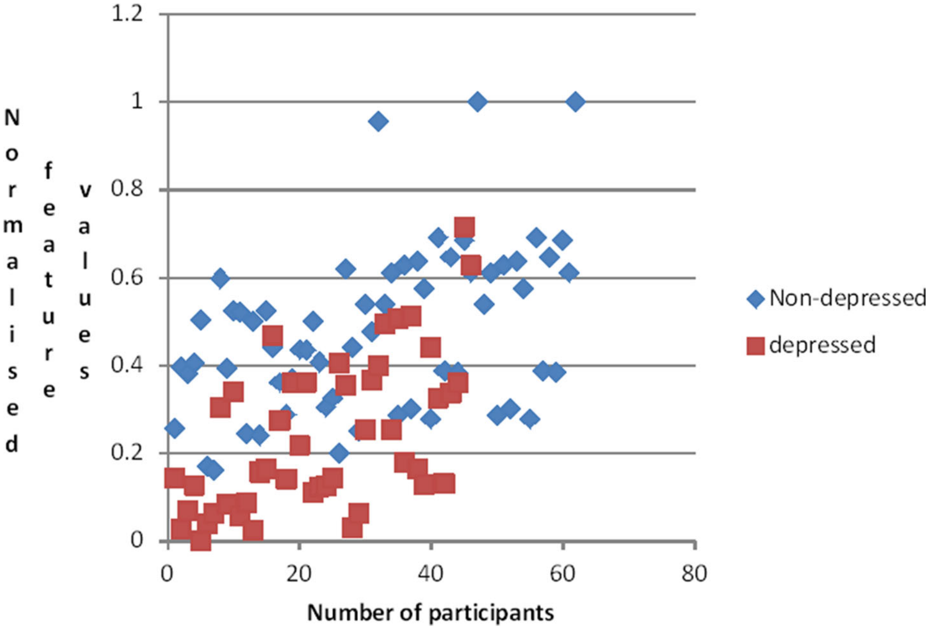


Fig. 10 Positive correlated single feature variation in all participants

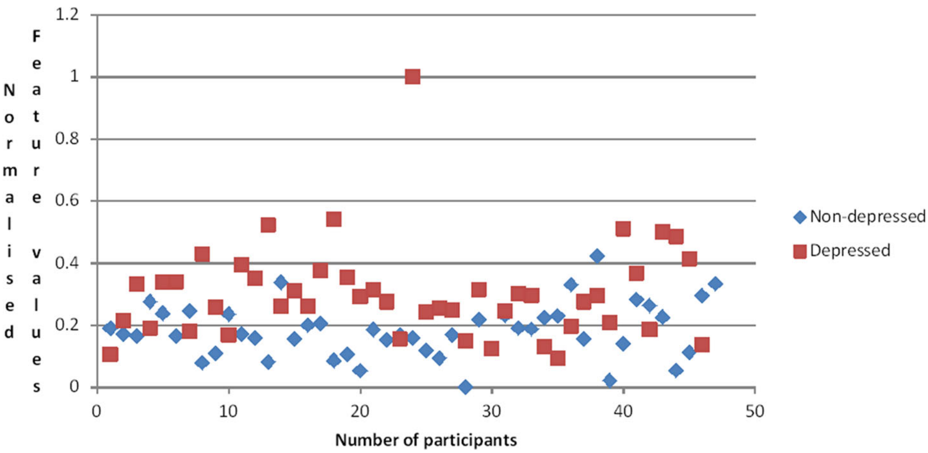
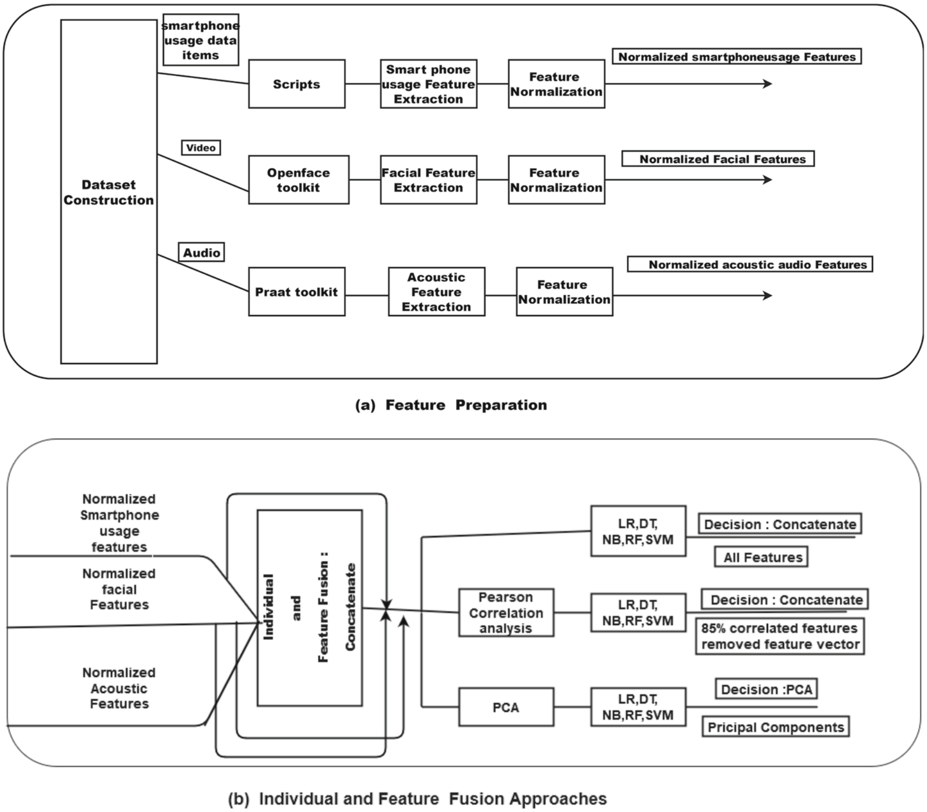


Fig. 11 Negative correlated single feature variation in all participants

Table 9 Average accuracy classification results for individual modalities

S.no.	Individual modalities	ML Classifiers	smart phone modality		visual modality		audio modality		Average Accuracy
			#	Acc	#	Acc	#	Acc	
1	All features	LR	53	61	415	70	82	60	64
2		DT		68		77		55	67
3		NB		62		78		68	69
4		RF		69		79		67	72
5		SVM		65		80		60	68
6	Pearson correlation reduced feature vector	LR	45	60	166	79	57	67	69
7		DT		50		80		60	63
8		NB		58		66		61	62
9		RF		50		80		68	66
10		SVM		55		80		72	69
11	PCA	LR	28-30	66	40-42	80	20-22	69	72
12		DT		68		69		52	63
13		NB		66		72		50	63
14		RF		66		79		50	65
15		SVM		69		80		69	73
Individual modality average				62		77		62	

#- Number of features in a resultant feature vector, Acc-accuracy, and average accuracy corresponds to the row average to demonstrate each ML classifier used in different methods. Individual modality average corresponds to the column average of the individual modality


smartphone usage data items
Dataset Construction
Scripts
Smart phone usage Feature Extraction
Feature Normalization
Normalized smartphoneusage Features
Video
Openface toolkit
Facial Feature Extraction
Feature Normalization
Normalized Facial Features
Audio
Praat toolkit
Acoustic Feature Extraction
Feature Normalization
Normalized acoustic audio Features
Normalized Smartphone usage features
Normalized facial Features
Normalized Acoustic Features
Individual and Feature Fusion : Concatenate
LR,DT, NB,RF,SVM
Decision : Concatenate
All Features
Pearson Correlation analysis
LR,DT, NB,RF,SVM
Decision : Concatenate
85% correlated features removed feature vector
PCA
LR,DT, NB,RF,SVM
Decision :PCA
Pricipal Components

(a) Feature Preparation

(b) Individual and Feature Fusion Approaches

Fig. 12 Summary of the investigated system configuration: a) feature preparation steps for smart phone usage, audio-visual modalities. b) using normalised feature vectors of different modalities:Individual and feature fusion techniques that were investigated

Table 10 lists the classification results using feature fusion(see Fig. 12b for visual representation):1)concatenating all the features of the individual modalities.(see row# 1 to 5). 2) concatenating the reduced feature vectors of individual modalities, which are obtained with Pearson correlation analysis(see row# 6 to 10), and 3) Applying PCA over concatenated feature vectors of individual modality(see row# 11 to 15).

From Table 9, it is evident that visual results are more encouraging than smart phone usage and audio modalities in all the performed ways. The reason could be among the features extracted, visual modality features show a higher correlation than smart phone and audio modality features. (refer Section 4.1).

From Table 10, using feature fusion, SVM with the concatenation based on Pearson's correlation analysis showed the best performance, i.e., 86% accuracy(see row# 10).where as naive byes and random forest showed slightly lesser accuracy rates of 83% and 85%, respectively(see row #8-9).

From Tables 9 and 10, the combination of modalities led to better performance in terms of accuracy. In most of cases, bi-modality performed better than uni-modality, tri-modality performed even better than bi-modality.

Table 10 Average Accuracy classification results for fused modalities

S.no.	Fused modalities	ML Classifiers	smart-phone+ audio modality		smartphone + video modality		video+ audio modality		Method Average	All modalities	
			#	Acc	#	Acc	#	Acc	Acc	#	Acc
1	Concatenate all features	LR	135	82	468	78	497	81	80	550	**83**
2		DT		81		78		80	80		80
3		NB		82		72		75	76		80
4		RF		79		83		83	82		83
5		SVM		81		79		83	81		**84**
6	Concatenate Pearson correlation removed feature vectors	LR	101	80	209	81	220	79	80	265	79
7		DT		81		79		82	81		80
8		NB		80		81		82	81		**83**
9		RF		85		80		80	82		**85**
10		SVM		83		83		84	83		**86**
11	95% of variance of PCA over concatenated feature vectors	LR	40-42	75	30-32	79	40-42	79	78	50-55	78
12		DT		70		62		60	78		65
13		NB		79		72		79	77		74
14		RF		83		73		65	74		**75**
15		SVM		81		79		80	80		**82**
Fused modalities average				80		77		78			**80**

#- Number of features in a resultant feature vector,Acc-Accuracy, and method average corresponds to the row average to demonstrate each ML classifier used in different methods. Fused modalities average corresponds to the column average of the fused modality. **Bold**: fused modalities performed well when compared with the method average. Fused modalities average is the column average to demonstrate the average of each modality combination

From the results, we hypothesise that high correlation among the features contributes to redundancy in the dataset. This could confuse the classifier. Therefore removing the redundant features will enhance the performance of the system.

4.3 Comparisons of results on benchmarking dataset

Lastly, we demonstrate the effectiveness of the proposed approach by the comparison of the results on a subset of feature vectors on DAIC dataset(widely used benchmarking dataset in depression detection). However, it is to be noted that this dataset contains only visual and audio data modalities and lacks smartphone usage data. To the best of our knowledge, we could not find any benchmarking dataset with all the three data modalities that we have proposed. Hence we have chosen this dataset to compare the results with the two data modalities.

The features (described in Section 3.4.2 visual feature extraction and Section 3.4.3 audio feature extraction) were extracted on the DAIC dataset, and the results were compared. Table 11 lists the average accuracies with the proposed feature vectors on ML Classifiers on DAIC dataset.

From Fig. 13, it is evident that LR and SVM using audio and video achieved 86% accuracy. Hence we believe that our approach can work on any kind of depressive diagnosis detection with similar cues.

Table 11 Results of proposed approach on DAIC Dataset

S.no.	ML Classifiers	Fused modalities	All features features		Pearson correlation reduced feature vector		PCA	
			#	Acc	#	Acc	#	Acc
1	Logistic Regression	Audio	82	70	50	81	20-25	68
2		Video	230	81	72	83	20-25	80
3		Video + Audio	312	83	122	**86**	30-35	83
4	Decision Tree	Audio	82	62	50	71	20-25	62
5		Video	230	80	72	80	20-25	82
6		Video + Audio	312	80	122	82	30-35	82
7	Naive Bayes	Audio	82	55	50	70	20-25	70
8		Video	230	80	72	75	20-25	80
9		Video + Audio	312	80	122	82	30-35	80
10	Random Forest	Audio	82	74	50	66	20-25	64
11		Video	230	85	72	80	20-25	81
12		Video + Audio	312	85	122	85	30-35	81
13	Support Vector Machines	Audio	82	74	50	68	20-25	67
14		Video	230	85	72	83	20-25	82
15		Video + Audio	312	85	122	**86**	30-35	83

# - number of features in the feature vector. Acc-Accuracy, and **BOLD**: Best accuracies obtained Note- DAIC dataset does not contain all the low-level openface feature sets. Hence we extracted statistical feature vector on the available low-level feature vector of DAIC dataset

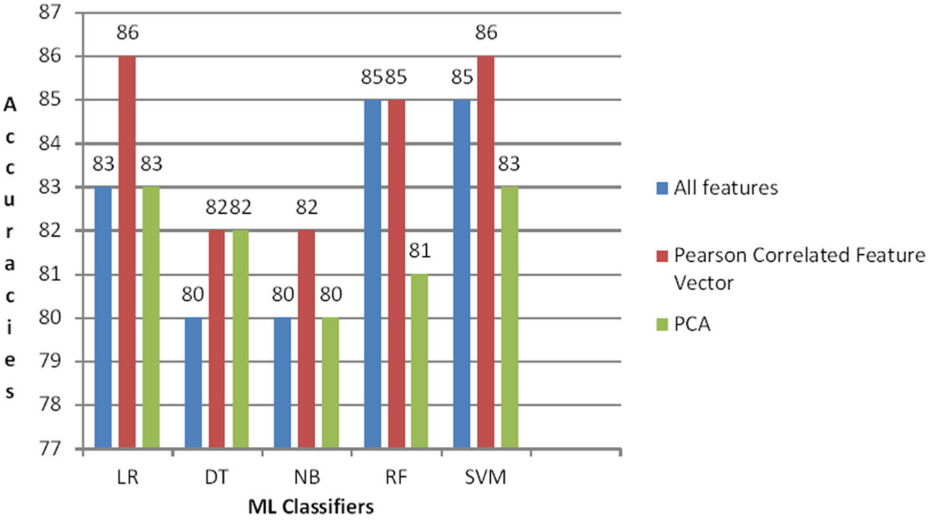


Fig. 13 Comparision of accuracies over ML classifiers with feature selection methods: All Features, Pearson's correlation analysis, and PCA using both audio and video

An alternative measure of accuracy,the Receiver Operating Characteristic(ROC) [29] is shown in Fig. 14. In ROC curve,a graph between true positive rate(sensitivity) and false positive rate(specificity) of ML classifiers were plotted. Curves were plotted using Pearson's

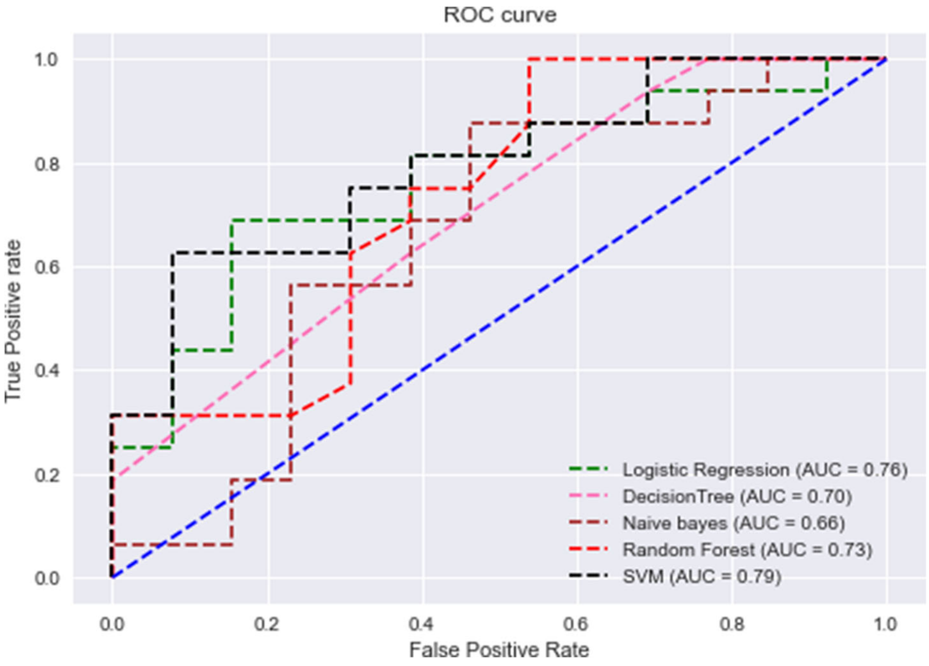


Fig. 14 ROC curve of ML classifiers

correlation reduced feature vector, which gave the best accuracy. The Area under the ROC Curve called as AUC(Area Under Curve) is also provided is Fig. 14. it can be seen that, the highest Score of AUC (79%) for SVM classifier that indicates better performance over other classifiers in classifying depressed and non-depressed subjects.

5 Conclusion

This study investigated multi modal features extracted from MCS and task/interview based mechanism to identify depressed and non-depressed participants. For this purpose, the user data was collected in a unique way by acquiring their smartphones usage data, emotion and speech elicitation mechanisms. In our research, we designed and experimented with an end-to-end machine learning approach, which involves multimodal data collection, feature extraction, feature selection, feature fusion, and classification to determine and distinguish depressed and non-depressed subjects. We experimented with: various features from multimodalities individually, and by fusing them, features selection techniques based on PCA and Pearson's correlation analysis, and different machine learning classifiers such as Logistic Regression, Decision Tree, Naive Bayes, Random Forest, Support Vector Machines for classification.

Our findings suggest that combining features from multiple modality performs better than any single data modality and the best classification accuracy is achieved when features from all the three data modality are fused. Also feature selection method based on Pearson's correlation coefficients improved the accuracy in comparison to using all the features and other selection technique like PCA. Amongst different machine learning classifiers that we experimented with, SVM yielded the best accuracy of 86%. Our proposed approach was also applied on a benchmarking dataset, and results demonstrated multimodal approach to be advantageous in performance with state-of-the-art depression recognition techniques.

6 Limitations and future work

It is a common problem in similar kind of studies, that the limited number of participants because of the selection criteria could affect the result analysis. A large-scale study using a clinically validated depression diagnosis is preferable. Demographic labels(gender, age, marital status, etc.), which are the main factors for depression diagnosis, can be further explored.

For the study, we have manually collected verbal and non-verbal cues using zoom meetings. Developing an automatic assessment technique using a mobile application that could monitor the mobile phone usage patterns and check the verbal and nonverbal indicators(when user provide permission) would be more beneficial.

In the present work, we have used a popular technique of Pearson's correlation for statistical analysis. However, different statistical analysis techniques and their comparative study could reveal the more scientific significance of the work.

Our proposed approach needs at least 14 days of smart phone usage patterns for classification, which could be a little costly. Any machine learning algorithm has to maintain a trade-off between the time duration required and output prediction. Future studies will build a classification approach with 3/7/10 days of the most influential factors that contribute to depression diagnosis.

Our future work will explore semantic cues in verbal, head pose and eye gaze features in visual, along with skin conductance and heartbeat in physiological modalities along with a focus on more advanced smart phone usage variables.

Acknowledgements The authors wish to thank all the participants who helped to do this research.

Declarations

Conflict of Interests There is no conflict of interest.

Ethics approval The experimental procedure used for the study is approved by the Institutional Committee of Visveswaraya National Institute of Technology, Nagpur, India. For this research study, volunteer's smart phone usage logs, visual and verbal data were used only after obtaining informed written consent forms.

References

1. Agarwal S (2013) Data mining: data mining concepts and techniques, 203–207 (IEEE)
2. Alghowinem S, Goecke R, Wagner M, Parker G, Breakspear M (2013) Eye movement analysis for depression detection, 4220–4224 (IEEE)
3. Alghowinem S, Goecke R, Wagner M, Parkerx G, Breakspear M (2013) Head pose and movement analysis as an indicator of depression, 283–288 (IEEE)
4. Alghowinem S et al (2016) Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. IEEE Trans Affect Comput 9(4):478–490
5. Asgari M, Shafran I, Sheeber LB (2014) Inferring clinical depression from speech and spoken utterances, 1–5 (IEEE)
6. Baltrusaitis T, Zadeh A, Lim YC, Morency L-P (2018) Openface 2.0: Facial behavior analysis toolkit 59–66 (IEEE
7. Barbosa PA, Madureira S (2016) Elicitation techniques for cross-linguistic research on professional and non-professional speaking styles, 503–507
8. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961) An inventory for measuring depression. Arch Gen Psychiatr 4(6):561–571
9. Ciman M, Wac K (2016) Individuals' stress assessment using human-smartphone interaction analysis. IEEE Trans Affect Comput 9(1):51–65
10. Colom F et al (2009) Group psychoeducation for stabilised bipolar disorders: 5-year outcome of a randomised clinical trial. Br J Psychiatry 194(3):260–265
11. Cootes T, Edwards G, Taylor C (2001) Robust real-time periodic motion detection, analysis, and applications. IEEE Trans Patt Analy Mach Intell 23(6):681–685
12. Cummins N, Epps J, Breakspear M, Goecke R (2011) An investigation of depressed speech detection: features and normalization
13. Cummins N, Sethu V, Epps J, Schnieder S, Krajewski J (2015) Analysis of acoustic space variability in speech affected by depression. Speech Comm 75:27–49
14. Cummins N et al (2015) A review of depression and suicide risk assessment using speech analysis. Speech Comm 71:10–49
15. De Vos M et al (2016) Detecting bipolar depression from geographic location data IEEE Transactions on Biomedical Engineering 64 (8)
16. Degottex G, Kane J, Drugman T, Raitio T, Scherer S (2014) Covarep—a collaborative voice analysis repository for speech technologies, 960–964 (IEEE)
17. Ekman P, Davidson RJ, Friesen WV (1990) The duchenne smile: emotional expression and brain physiology: Ii. J Pers Soc Psychol 58(2):342
18. Ekman P, Matsumoto D, Friesen WV (1997) Facial expression in affective disorders. What the Face Reveals: Basic and Applied studies of Spontaneous Expression Using the Facial Action Coding System (FACS) 2:331–342
19. Eyben F, Wöllmer M, Schuller B (2010) Opensmile: the munich versatile and fast open-source audio feature extractor, 1459–1462
20. Fukazawa Y et al (2019) Predicting anxiety state using smartphone-based passive sensing. J Biomed Inform 93:103151

21. Fukazawa Y et al (2020) Smartphone-based mental state estimation: a survey from a machine learning perspective. J Inf Process 28:16–30
22. Girard JM, Cohn JF, Mahoor MH, Mavadati S, Rosenwald DP (2013) Social risk depression: evidence from manual and automatic facial expression analysis, 1–8 (IEEE)
23. Gratch J et al (2014) The distress analysis interview corpus of human and computer interviews, 3123–3128
24. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC (2017) Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci 18:43–49
25. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis, 46–53 (IEEE)
26. Kanter JW et al (2003) Comparison of 3 depression screening methods and provider referral in a veterans affairs primary care clinic. Prim Care Comp J Clin Physchiatry 5(6):245
27. Kelly D, Condell J, Curran K, Caulfield B (2020) A multimodal smartphone sensor system for behaviour measurement and health status inference. Information Fusion 53:43–54
28. Kelly D, Curran K, Caulfield B (2017) Automatic prediction of health status using smartphone-derived behavior profiles. IEEE Biomed Health Inform 21(6):1750–1760
29. Khan A, Zubair S (2020) An improved multi-modal based machine learning approach for the prognosis of alzheimer's disease. Journal of King Saud University-Computer and Information Sciences
30. Kroenke K, Spitzer RL, Williams JB (2001) The phq-9: validity of a brief depression severity measure. J Gen Intern Med 16(9):606–613
31. Kumar S, Chong I (2018) Correlation analysis to identify the effective data in machine learning: prediction of depressive disorder and emotion states. Intern Env Res Public Health 15(12):2907
32. Latif S et al (2020) Leveraging data science to combat covid-19: a comprehensive review. IEEE Trans Artif Intell 1(1):85–103
33. Li M et al (2020) Method of depression classification based on behavioral and physiological signals of eye movement. Complexity 2020
34. Lin LY et al (2016) Association between social media use and depression among us young adults. Depression and Anxiety 33(4):323–331
35. Littlewort G et al (2011) The computer expression recognition toolbox (cert) 298–305 (IEEE)
36. Liu J, Shen H, Narman HS, Chung W, Lin ZA (2018) Survey of mobile crowdsensing techniques: a critical component for the internet of things. ACM Trans Cyber-Phys Syst 2(3):1–26
37. Masud MT et al (2020) Unobtrusive monitoring of behavior and movement patterns to detect clinical depression severity level via smartphone. J Biomed Inform 103:103371
38. Morales M, Scherer S, Levitan R (2017) A cross-modal review of indicators for depression detection systems, 1–12
39. Moshe I et al (2021) Predicting symptoms of depression and anxiety using smartphone and wearable data. Frontiers in psychiatry 12
40. Moshe I et al (2021) Predicting symptoms of depression and anxiety using smartphone and wearable data. Frontiers in psychiatry 43
41. Narziev N et al (2020) Stdd: short-term depression detection with passive sensing. Sensors 20(5):1396
42. Nasir M, Jati A, Shivakumar PG, Nallan Chakravarthula S, Georgiou P (2016) Multimodal and multiresolution depression detection from speech and facial landmark features, 43–50
43. Organization WH et al (2017) Depression and other common mental disorders: global health estimates. Tech. Rep. World Health Organization
44. Pabba C, Kumar P (2022) An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. Expert Syst 39(1):e12839
45. Pampouchidou A (2018) Automatic detection of visual cues associated to depression. Ph.d. thesis schooluniversité Bourgogne franche-comté
46. Pampouchidou A et al (2016) Depression assessment by fusing high and low level features from audio, video, and text, 27–34
47. Pampouchidou A et al (2017) Automatic assessment of depression based on visual cues: A systematic review. IEEE Trans Affect Comput 10(4):445–470
48. Panicker SS, Gayathri P (2019) A survey of machine learning techniques in physiology based mental stress detection systems. Biocybern Biomed Eng 39(2):444–469
49. Pediaditis M et al (2015) Extraction of facial features as indicators of stress and anxiety, 3711–3714 (IEEE)
50. Ray A, Kumar S, Reddy R, Mukherjee P, Garg R (2019) Multi-level attention network using text, audio and video for depression prediction, 81–88
51. Rush AJ et al (2003) The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 54(5):573–583

52. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC (2016) The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 4:e2537
53. Salari N et al (2020) Prevalence of stress, anxiety, depression among the general population during the covid-19 pandemic: a systematic review and meta-analysis. Glob Health 16(1):1–11
54. Scherer S, Lucas GM, Gratch J, Rizzo AS, Morency L-P (2015) Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews. IEEE Trans Affect Comput 7(1):59–73
55. Seppälä J et al (2019) Mobile phone and wearable sensor-based mhealth approaches for psychiatric disorders and symptoms: systematic review. JMIR Mental Health 6(2):e9819
56. Stasak B (2018) An investigation of acoustic, linguistic and affect based methods for speech depression assessment
57. Uhrig MK et al (2016) Emotion elicitation: A comparison of pictures and films. Front Psychol 7:180
58. Wang Q, Yang H, Yu Y (2018) Facial expression video analysis for depression detection in chinese patients. J Vis Commun Image Represent 57:228–233
59. Williamson JR, Quatieri TF, Helfer BS, Ciccarelli G, Mehta DD (2014) Vocal and facial biomarkers of depression based on motor incoordination and timing, 65–72
60. Williamson JR et al (2016) Detecting depression using vocal, facial and semantic communication cues, 11–18
61. Xiong H, Huang Y, Barnes LE, Gerber MS (2016) Sensus: a cross-platform, general-purpose system for mobile crowdsensing in human-subject studies, 415–426

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.