Measurement: Sensors 25 (2023) 100587

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

A hybrid model for depression detection using deep learning

Vandana [a], Nikhil Marriwala [b,*], Deepti Chaudhary [b]

[a] University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, India
[b] Electronics and Communication Engineering Department, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, India

ARTICLE INFO

Keywords:
Deep learning
Convolution neural network
Textual CNN
Audio CNN
LSTM
Bi-LSTM
Training loss
Validation loss
Training accuracy
Validation accuracy
Confusion matrix

ABSTRACT

Millions of people are suffering from mental illness due to unavailability of early treatment and services for depression detection. It is the major reason for anxiety disorder, bipolar disorder, sleeping disorder, depression and sometimes it may lead to self-harm and suicide. Thus, it is a very challenging task to recognize people who are suffering from mental health disorders and provide them treatments as early as possible. Conventionally, depression detection was done through patients' interviews and PHQ scores, accuracy of conventional methods is very less. In this work, a hybrid model is proposed for depression detection using deep learning algorithms, which mainly combines textual features and audio features of patient's responses. To study behavioral characteristics of depressed patient's, DAIC-WoZ database is used. Proposed method consists of three components; first, a textual CNN model in which a CNN model is trained with only text features, second, an audio CNN model in which CNN model is trained with only audio features and third, a combination of audio and textual model named as hybrid model in which LSTM algorithms are applied. An improved version of LSTM model named as Bi-LSTM model is also used in the proposed work. In results, training accuracy, training loss, validation accuracy and validation loss is calculated for all the mentioned models. The results shows that deep learning is a better solution for depression detection in which accuracy of textual CNN model is 92% whereas accuracy of audio CNN model is 98% and loss of textual CNN is 0.2 whereas loss of audio CNN is 0.1. These results show that audio CNN is a good model for depression detection. It performs better as compared to textual CNN model. It is also observed that Bi-LSTM has better learning rate as compared to other models with accuracy 88% and validation accuracy 78%. There are some parameters such as precision, F1-score, recall and support are found for evaluation of models. In results, graphs for training loss, validation loss, training accuracy and validation accuracy are plotted. At last, by using confusion matrix depression can be detected for textual CNN Model, audio CNN model, LSTM model and Bi-LSTM against true label and predicted label.

1. Introduction

Depression is a medical condition and it is one of the most common mental illness which affects millions of people globally. Depression is considered as a dangerous disease, which not only affects mental state of a person but also cause physical harm to a patient. Severity of Depression is predicted in terms of mental health condition of a patient [1]. Most frequent cases of Mental Health Disorders include anxiety disorder, restlessness, sleeping disorder, eating disorder, addictive disorder, Depression, Trauma, and stress related disorders [2]. Depression is a type of mental illness in which a patient continuously feels hopelessness, demotivation, bad mood swings and loss of interest in daily physical, mental and social activities which leads to emotional damage and physical changes in a patient's body. It specially affects learning capacity in a person, causes mood fluctuations and it often reduces work efficiency in a person. A depressed patient has different symptoms based on its severity [3]. In high severity, brain activity goes slow down and produces the hormone cortisol which affects the growth of neurons in brain. It badly affects the thought process of a person and sometimes it may lead to suicidal cases. Depression has some stages of occurrence, clinical depression, bipolar depression, dysthymia, seasonal affective disorder and others [4]. There are some treatment options that are available, treatments and services vary from counseling sessions to therapies. Brain simulation treatments are also available [5]. According to a survey conducted by WHO, approximately 280 million people are suffering from depression worldwide, nearly 800000 cases of suicide due to depression reported every year worldwide [1]. Depression is the fourth leading disease in this world, it became a leading problem which

* Corresponding author.
E-mail addresses: vandana301098@gmail.com (Vandana), nmarriwala@kuk.ac.in (N. Marriwala), deeptic2015@kuk.ac.in (D. Chaudhary).

https://doi.org/10.1016/j.measen.2022.100587
Received 30 September 2022; Received in revised form 7 November 2022; Accepted 20 November 2022
Available online 30 December 2022

affects people of all age groups children, teenage, adults, old age people. More than 80% people not get proper treatment due to unavailability of early services and treatments for depressed patients [6,7]. Researchers estimated that around 1 out of 5-person experience depression once in their lifetime. As number of people suffering from depression increases day by day, it is very important to see the severity of this mental condition. Therefore, an automatic system is required to detect early signs of depression and treat patients accordingly. Most used tools for depression detection are interview style system and assessment system such as the Hamilton Rating Scale for Depression and self-reporting techniques such as Beck Depression Inventory tool and PHQ-8 score, Structured Clinical Interviews are designed to predict symptoms severity and common actions observed in depression patients [8–10]. Traditional therapies and medications for depression such as psychotherapy or pharmacological services are mostly time consuming, expensive, and ineffective. Major problem with these traditional methods is that, firstly these depression detection techniques need more patient data, their background, their history, and any past trauma related information, to predict symptoms of depression or these treatments need continuous monitoring on patient activities for prediction of depressed patient or not [11]. And secondly, fear from public or society brings other negative consequences which may affect the diagnosis, patients intentionally hide their real response and conditions from doctors intentionally due to pressure from society and then they often mislead the treatment which consume more time for diagnosis [12]. Research shows that a depressed patient is different from a normal person due to changes in thought process, facial movements, body movements, facial expressions, mental behavior, physiological signals, and psychological signals [12]. A depressed patient often stammers while talking and take uneven pauses in speech. Another symptom is incorrect pronunciation; a depressed patient pronounces words and sentences in a slow and incorrect manner [13]. Research shows that the response time is significantly longer in case of depressed patient. They take more time to listen, respond and action which show symptoms of depression [6,8]. Individuals with depression tendencies are disturbed by sad thoughts and show a noticeable bias towards negative stimuli. Depressed patient continuously uses words that contain words with rejection with negative expressions which indicate sadness, stress, demotivation, or dissatisfaction [14]. An irregular menstrual cycle in females also indicates symptom of stress and depression [15]. Other factors such as abnormal eye contact, less frequent mouth movement, less activity detected in patients dealing with depression disorder [16,17]. People suffering from depression disorder have different behaviors on psychological level than normal people. For example, the generated brain signal and the level of feeling good hormones such as serotonin and oxytocin in the body are different [18]. EEG signals (Electroencephalography), NMR (nuclear magnetic resonance) signals and other audio or image signals are different in case of depressed patient [19,20]. As a result, psychologists cannot accurately predict symptoms of depression and its severity, which makes conditions of depressed patient worst. Therefore, an accurate, automatic, and accessible approach is a necessity for patients as well as for psychologists.

1.1. Related work

KBRS is developed which contains a system for monitoring of emotional health, it uses deep learning model and metric of sentiment named eSM2. This system identifies which sentences present negative content using ML and DL algorithms. It uses CNN and Bi-LSTM algorithm and detect with the accuracy of 0.89 and 0.90 to detect depressed patients and stressed patients [6]. Also, this monitoring system is able to send warning texts to people who are either depressed or stressed. Some studies use social media data like Facebook, Instagram and reddit posts to analyze patient behavior on social media platforms. Through social media posts, we can easily find behavior of a patient either depressed or stressed or other mental health disorders. The method is employed through Co-training technique (type of semi-supervised learning approach) by incorporating the discriminative power of widely used classifiers namely RF and SVM and NB with an accuracy of 0.70 and 0.85 [3]. A solution proposed in which a hierarchical posts representation model named MGL-CNN model for identifying depressed people in social media platforms [8,21]. This method consists of operations in post level and operations in user level. Besides, another depression detection model named as SGL-CNN also developed by changing number of gates in model architecture. This model predicts Precision, Recall, F1-score values [22]. The study mainly consists of approaches used for detection of depression by using NLP and text classification techniques. It mainly recognizes lexicon terms which are common to depressed patients. Results of this study helps to improve accuracy and performance of model [9,19]. The best feature in the research is bigram with the classifier used is SVM which helps in detection of depression with 80% accuracy and 0.81 F1-score. Some research uses DCNN and ANN for depression recognition. Two models are proposed named as Deep model and Shallow model for analysis of depressive symptoms. This paper proposes a text and video features and combines deep and shallow models [2]. It consists of RF algorithm for depression classification from score estimation. A new depression detection method is proposed which not only utilize speech signals but also extract text based on language from patient responses [4]. The proposed model mainly contributes research for detection of mental illness which includes Bi-LSTM network with attention layer to recognize text data and, a 1-D CNN to recognize audio signals and a fully connected network which concatenate the outputs of previous two models to assess the severity of depressive state [23]. In research paper [7], which uses text data from university students to detect depression among university students. DISVM algorithm is utilized to classify data obtained from input and finally recognize depression mental illness. According to results in Ref. [7], precision of DISVM is 0.88 on training set and 0.86 on testing set. Some studies use multi-modal data from text, audio and video and then predict an output regarding the mental illness of patient [5]. The output is divided into different levels of depression to take into consideration the severity of depression of the patient [8]. These are some of the approaches which are used for depression recognition, and classification of text, video, and audio features. According to results in Ref. [5], f1-score is 0.81, precision is 0.80. In this way, some models predict depression symptoms and its severity accurately while others not give satisfactory results. From above literature work, researchers concluded that there are many solutions proposed by scientists to the problem of depression detection [6]. As increasing number of cases of depression leads to many proposed solutions but these solutions not achieved high accuracy yet and losses are high in proposed solutions [24]. There are some researchers use data from social media platforms, which may be accurate or not [25]. As prediction of depression from using online platforms such as Twitter and Reddit may cause a risk that these evidences can wrongly predict symptoms of depression. If symptoms observed from online social media platforms will wrongly predicted, this will create a blunder [13]. If the results are not accurate, then certain risk of depression cannot be evaluated. Sometimes in social media platforms, users may intentionally or unintentionally post depressed or sad stories which will affect entire process of depression detection system [26]. Therefore, researchers cannot properly depend on social media platforms. They tend to use some accurate database from developing a system for depression detection. Researchers need an automatic depression detection system which achieved high accuracy and minimum losses evaluated from the system. Concatenating the features of audio samples, video samples and text responses of a depressed patient can give accurate results [27]. Using Deep Neural Network, system can easily predict depression [28,29]. When model is trained and learn all the features of audio, video, and text, only then depression detection system will be developed. These are Literature Survey related to Depression detection problem.

2. Proposed model

Proposed Model is based on two algorithms i.e., one is CNN and second is LSTM. CNN or Convolution neural network or ConvNet is a deep learning algorithm which takes input as an image, assigns some weights to input values and then helps in classification of output image [30]. CNN not only used in image classification but also for data analysis, pattern recognition, and computer vision and solving NLP tasks [8]. CNN is a class of Artificial Neural Network (ANN), it is also known as Multilayer Perceptron. CNN is an algorithm which is inspired from design of neurons in human brain [8]. CNN works on the principal of convolution operation. Convolution Operation is performed as shown in below example and shown in Fig. 1 as follows. For example- Input score * Kernel/filter (same size) = Output score (Feature Map) [10]. Advantage of using CNN Algorithm: requirement of processing time in CNN is much lower as compared to other algorithms. Applications of CNN Algorithm is that, it is mostly used in visual images, speech recognition, pattern recognition, text recognition and understanding natural language processing problems. Architecture of CNN algorithm deal with no. of convolution layers, max-pooling layers, and fully connected layers. ReLU as an activation function is used in proposed work. To remove problem of over fitting, Dropout layers are also used with value of 0.25. Flatten layers are used along with convolution layers [11,31]. Output layer is in the form of Binary labels: Depressed and Not-Depressed (0 and 1). Architecture of CNN Architecture is shown in Fig. 2 as.

To begin proposed work with text classification in textual CNN, work embedding layer and CNN layer is used. Basically, word embeddings are vector or picture representation of words. Word2vec meaning word to vector is the most popular technique of word embeddings. Word2vec is obtained using Skip Gram and CBOW model. Word2vec is a two-layer model that process text data [31]. Input of Word2vec is text data and output is vectors or pictures. Vectors are features vectors for words in the collection. The purpose of using word2vec is to group the vectors of similar words together in vector space. It detects similarities mathematically [32]. Word2vec creates vectors which composed of numerical representation of word features, features such as the data or information of individual words. Word2vec develop this feature without human intervention. Word2vec make predictions of high accuracy about meaning of a word based on previous experiences [33]. Those predictions used to obtain similarities in a word with another words. Word2vec give training to other words which share similar neighbor with input data. This can be done in one of two ways, either using data to predict a word which is target, a method known as CBOW or using a word to predict a data in target, which is called Skip Gram. From above two ways, both have their own set of advantages and disadvantages. Skip Gram works well only with less amount of dataset whereas CBOW works faster and has better learning rate. For proposed work, GoogleNews-vectors-negative300.bin is used [32]. It is a word2vec model developed by google. It is a pretrained model which contains set of words that will be used for text classification. It works well on sentiment analysis. It can be downloaded from Kaggle website. CNN works well with image recognition and pattern recognition. For text

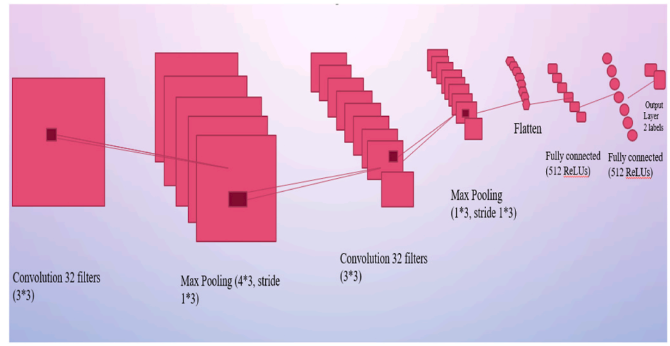


Fig. 2. CNN architecture.

classification purpose, only CNN model not used [34]. Along with CNN mode, a word2vec is also used. Dataset for text classification for the purpose of depression detection is in the form of text responses of depressed patients. For example- GoogleNews-vectors-negative300.bin. Results of text classification is in the form of Binary labels [35]. Similarly, for audio recognition in audio CNN, spectrograms are collected from audio samples. Audio samples are in the form of audio recordings of depressed patients. Audio samples converted into spectrograms [36]. Audio features such as pitch tone, rhythm, stress, voice quality, articulation, intonation, Mel spectrograms, MFCC [37]. There are some steps for audio classification which includes: segmentation, data cleaning, feature extraction, data imbalance [12]. Depression detection from audio samples is a challenging task. The first step for audio classification is to convert an audio sample into spectrogram [38]. This is important step for audio classification. A Spectrogram is a visual representation of frequencies of a signal as it varies with respect to time. After conversion of audio sample into spectrograms, next step is Audio Splitting [39]. In audio splitting, removal of extra noise and silence from audio samples, this step is also called Segmentation. After the removal of undesired noise and silence from audio or speech sample, next step is Data Imbalance. In dataset, Information of non-depressed person is more than that of depressed patients. It is four times more than data of depressed patients. That's why Data Imbalance is important. Balance the data of Depressed: Non-Depressed into equal numbers [14]. Third step is Spectrogram Conversion. The sampled audio segments are then converted into spectrograms images of size 512*512 pixels. These images are put into training and validation folders in the ratio of 8:2. After this, Image Processing can be done. CNN algorithm can be applied on those images and predictions can be done for depressed and non-depressed patients. Results of audio classification are in the form of Binary Labels.

LSTM or (Long Short-Term Memory) Algorithm is a Recurrent Neural Network (RNN) in which mostly features linked with a layer to previous layer, it also allows information to pass from past to present and then Present to Future. RNN operate over sequence of vectors. Thus, each layer depends on previous outputs. Problem with RNN is that information rapidly gets lost with the passage of time [16]. LSTM are special

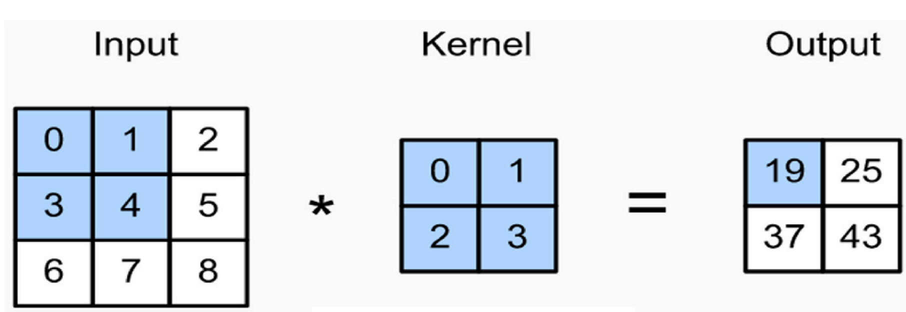


Fig. 1. Convolution Operation [2].

kind of RNN, they are designed to solve the problem of loss of information in RNN. LSTM are capable of learning long time dependencies which make RNN smart enough at remembering things. Advantage of using LSTM is, it will help in data processing predictions and pre-processing applications. Traditional algorithms like SVM, RF suffer from gradient vanishing problem [20]. LSTM improves model performance by memorizing important data. Disadvantage of using LSTM is that it requires more computational time for training of model [21]. More training time will make the model worse in some problems of Machine Learning [35]. Therefore, we use LSTM with minimum layers so that less training time will be required. LSTM is an algorithm which deals with Long Term and Short-Term information [22,25]. It has Previous Cell state, previous hidden state, new cell state, new hidden state, input data and output data. There are three gates in LSTM which are: forget gate, input gate and output gate in LSTM. These gates work as filters and data process in gates sequentially [36]. LSTM work on feedback mechanisms which make model to learn parameters effectively. They learn from past information and update past data to new data [18]. The entire sequence of data trained sequentially because of the availability of feedback mechanism in LSTM [40]. They remove vanishing gradient problem from the model and make the model works better as compared to other algorithms [41,42]. There are some layers in LSTM model which are sequential layer, LSTM layer, fully connected layer, SoftMax and output classification layer [11,43]. Bi-LSTM or Bi-directional LSTM is the process of training a neural network to make predictions from both directions i.e., future to past and past to future [11,29]. In Bi-LSTM, inputs have to flow in both directions which makes model better at remembering things from past and update new information in present [44]. This model can be used for text classification, speech recognition and pattern recognition [34,45]. This model work in both directions forward and backward, that's why this model is smart enough at predicting classification of sequential data [28,46,47]. Architecture of LSTM model is shown in Fig. 3 as.

2.1. Dataset

DAIC-WOZ Depression Database is used for Automatic Depression Detection system. Meaning of DAIC-WOZ is Distress Analysis Interview Corpus-Wizard of Oz Interviews [48]. The database is a collection of large corpus; it includes responses from depressed patients. This database is proposed to support the treatment of mental health disorders such as anxiety, depression, and stress disorder. It records information from patients in the form of audio or voice recordings, video recordings and questionnaire text responses. Therefore, this data is used due to availability of large amount of data in the form of audio, video and text responses. DAIC-WOZ Database downloaded from USC University of South California website. First, we have to apply on their website and then they will provide access to download the dataset. The dataset contains 189 sessions in the form of zip files, from 300_P.zip to 492_P.zip [49]. The above dataset contains responses and questionnaires from 59 depressed and 130 non-depressed people. This Database contains verbal and non-verbal symptoms of depressed patients in the form of audio features, video features and text features. The DAIC-WOZ database has been split into 3 folders i.e., dev_data, train_data, test_data. In each of these folders some files are present which are AUDIO.wav, COVAREP.csv, TRANSCRIPT.csv, FORMANT.csv etc [49].

- The training set file is in.csv format. This file consists of IDs of patients, Patient PHQ8 scores, labels in Binary form, gender of patient, and answers for every question of the questionnaire for training dataset.
- The development or dev set file contains Ids of patients, Patients PHQ8 scores, gender of participants, labels in binary form and answer for every question of the questionnaire for development dataset.
- At last, test dataset contains IDs of patients and patient gender information.

2.2. Software

Google Collab is the software that is used for depression detection in patients. It is an online platform, an open-source platform which is easy to use. Collab means Collaboratory which allows anybody to execute their respective codes in python language. Google collab is an online platform which runs in google server, it is in cloud and it connects the code with google drive in which all our database is stored. It gives access to GPU and TPU which automatically helps in fast processing of neural network and deep learning algorithms. For using free collab notebooks, a google account is needed which automatically connects to google drive. Therefore, no additional setup, installations and downloads are required which give user a friendly interface to perform coding and executions. Total free RAM available for processing is 16 GB. We can extend it to 25 GB and more, by upgrading Google Collab to Google collab pro. Maximum runtime connection is 12 h. After completion of 12 h, runtime will automatically disconnect from collab notebook. There is a variable panel present in google collab which variables, their values and their data types can be observed and evaluated. Kernel menu is absent in google collab, one kernel is present at one time and it can be re-evaluated using restarting the runtime. To work in deeper layers, TPU is present in Google collab. Most important thing is that with the help of

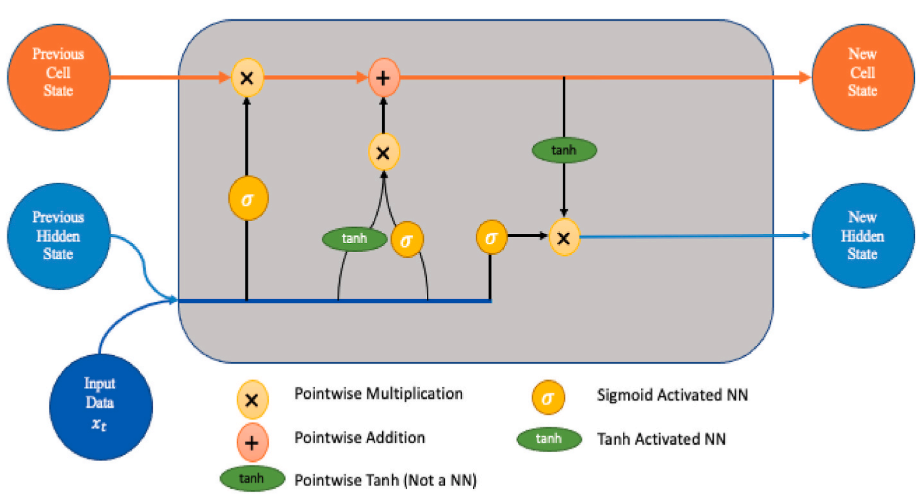


Fig. 3. LSTM architecture [12].

Google collab, notebooks can be shared very easily with other team members with the feature of google collab. Additional changes can be done very easily and frequently in google collab notebooks. Another important advantage of using Google collab is that it performs in less computational power as compared to Jupiter notebook which runs on system RAM. Proposed work is connected with the deep learning algorithms, for past processing GPU is used. With the help of Google Collab platform, Depression Detection using CNN and Bi-LSTM algorithm becomes easy.

2.3. Architecture of proposed work

Architecture of proposed model is divided into three components; first is textual CNN model which contain only text features, model is trained for only text features and second is audio CNN model which only contain audio features like Mel spectrograms, MFCC's, model is trained with only audio features. And third is a combination of above two models i.e., audio model + textual model = hybrid model [27]. This architecture mainly consists of some layers which are input layer, convolution layer, pooling layer, fully connected layer, SoftMax layer and output layer [50,51]. A neural network is designed when these layers start making connection with each other. In addition, some other layers such as dropout layer, activation function, flatten layer, batch normalization are also necessary in designing CNN architecture [52]. Some of the audio features which are used in proposed model are MFCC, COVAREP and Mel Spectrograms. For text features, pretrained model is used which already contain features of text responses of depressed patients. When compared with all features for audio CNN model, Mel Spectrograms produce best results. In a proposed model, audio features and text features convolve with CNN and LSTM architecture and generate results in binary labels: Depressed or Not-Depressed. For depression detection, proposed model is used, it automatically predicts whether a person is Depressed or not. The framework of the proposed method is shown in Fig. 4 as.

2.4. Layers

1) Convolution Layer- Convolution layer is the first layer and is the most important layer of a neural network. This layer creates a building block of an entire neural network. The main purpose of Convolution layer is to detect type of input i.e., text features, audio features or both. In convolution layer, input image is convolved with filter of same size and then output image is obtained. In the output a feature map is developed. Filter kernels are used in convolution layer which are weights, weights are updated by using back propagation algorithm in convolution layer.
2) Max-pooling Layer- The pooling layer is in between two convolution layers. Primary work of a pooling layer is to decrease size of input. This layer helps the model to reduce actual size of data and update the data with only necessary information of data. This layer improves efficiency of a neural network. It reduces undesired characteristics from data. The pooling layer creates a bridge between convolution layer and fully connected layer. Pooling layers also help to reduce over fitting problems happened in neural network.
3) ReLU as an Activation Function- ReLU means Rectified Linear Unit, this function describes non-linearity. This layer replaces all negative data with zero values. This layer is important layer because it decides which information to pass on to next convolution layer and which information to discard. There are many activation functions such as SoftMax, ReLU, tanH, Sigmoid.
4) Fully Connected Layer- A Fully Connected layer generally placed before output classification layer. In this layer, classification begins to happen and results update themselves. In a neural network, one or two fully connected layers are important to use.
5) Batch Normalization- Batch Normalization layers normalize the output of previous layers. It helps model to learn features effectively. It makes model stable and perform faster. It makes processing and learning of model faster.
6) Dropout layer- Dropout is a layer which helps to reduce over fitting in model. It randomly drops some values of neural network and makes a model learning rate faster. Dropout layer is placed after fully connected layer. Recommended value for dropout is 0.25. It can cause slow training in neural network.

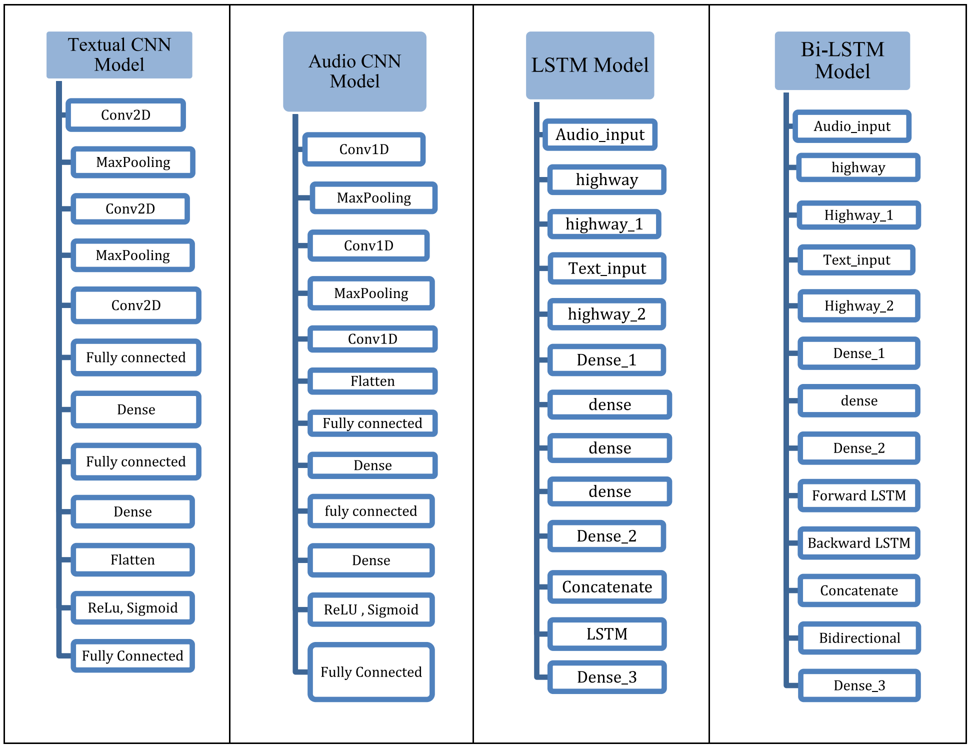
3. Experimental results

In experimental results, three models are obtained i.e., first is textual CNN Model, second is audio CNN Model and third is a hybrid model which is a combination of audio and text features. Some algorithms such as LSTM and Bi-LSTM are applied on hybrid model. Therefore, two hybrid models i.e., hybrid LSTM and hybrid Bi-LSTM model are obtained from proposed model. Adam Optimizer is used in all of the above models. With the help of these trained models for the purpose of depression detection, predictions can be made with achievement of high accuracy and minimum loss in proposed work. To predict performance of models, graph is plotted for all models. Graph of training loss and validation loss against no. of epochs and training accuracy and validation accuracy against no. of epochs is plotted. The textual CNN model architecture basically composed of mixture of convolution layers, fully connected layers, max pooling layer, flatten layer, dropout layer, activation function and loss function. The idea of designing of textual CNN model is to connect three convolution layers with three fully connected layers and some other layers such as activation function, max-pooling layer, flatten layer and dropout layer are added. Second model is audio model in which connection of three layers of convolution with three fully connected layers and other layers such as max-pooling flatten, ReLU as activation function are added in audio model. And third model is a hybrid model which is a combination of textual model and audio model. LSTM layers are added and concatenation is added for hybrid LSTM model. More convolution layers are included and hence more Max-pooling layers are added to make model a hybrid model. For Bi-LSTM model, forward LSTM and backward LSTM layers are added. Other layers such as flatten, activation function; concatenate layer is added in hybrid model. The layers of textual CNN model, audio CNN model, hybrid LSTM model and hybrid Bi-LSTM model are shown in Table 1.

According to experimental results obtained from audio CNN, textual CNN, LSTM and Bi-LSTM model, textual CNN model which only includes text features, model gave accuracy after training is 0.92% for 10 number of epochs with a loss of 0.3 and the execution time after training is very less. The execution time for textual CNN model is 1 min 42 s. For second model i.e., audio CNN model, the model gave accuracy after training is 0.98% for 10 number of epochs with a loss of 0.1 and the execution time after training is more as compared to textual CNN model. The execution

Fig. 4. Architecture of proposed Model.

Table 1
Layers in Proposed Model.



time for audio CNN model is 5 min 30 s. A hybrid LSTM model gave accuracy of 0.80% with a loss of 0.4. Hybrid LSTM model includes features of textual CNN model and audio CNN model and then LSTM algorithm is applied on trained model. Time required to train hybrid LSTM model is 2 h 26 min. At last, hybrid Bi-LSTM model is trained with audio and text features. Accuracy obtained by Bi-LSTM model is 0.88 which is more than Hybrid LSTM model and loss is 0.2 which is less than hybrid LSTM model. This means for depression detection, Bi-LSTM model predict accurately than LSTM model. Time required for training of Bi-LSTM model is 5 h 44 min. Time of training is more as compared to LSTM model. Table 2 shows comparison between models and evaluates accuracy, loss, val_accuracy, val_loss. It shows that training accuracy is more than validation accuracy and training loss is less than validation loss. Table 3 shows that evaluation parameters such as precision, recall, F1-score, support and these parameters define the performance of models. All models achieve good value of precision as shown in Table 3. As compared with [4], proposed model give better results and performance of proposed model is better. Precision values are better as compared to Ref. [4], F1-score for all models is better, support is same for all models, By using confusion matrix, severity of depression can be predicted. Confusion Matrix is in the form of scores or numbers between True Label and Predicted Label, these scores will predict a patient is depressed or not-depressed. This matrix is used for classification purpose. This matrix describes about the model is tend to be confused for predicting different classes. There are 4 labels for 2 classes. For Example- TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative). For Depression Detection, two classes are formed i.e., one is depressed while other is not depressed. These classes are labeled as 0 and 1. Confusion Matrix is shown in Table 5 as. As shown in Confusion matrix, Bi-LSTM predicts more number of depressed patients as compare to other models. This means, Bi-LSTM model is accurately predicted by Confusion Matrix (see Table 4).

A graph of accuracy, loss, val_accuracy and val_loss is plotted for all models as shown in Fig. 5, Fig. 6, Figs. 7 and 8 to compare performance of textual CNN model, audio CNN model, hybrid LSTM model, hybrid Bi-LSTM model [3]. In Figs. 5 (a), 6(a) 7(a), 8(a), a graph is plotted between

Table 2
Evaluation parameters accuracy, loss, Val_Accuracy, Val_Loss

METHOD	ACCURACY	LOSS	VAL_ACCURACY	VAL_LOSS
TEXT CNN	0.92	0.3	0.80	0.5
AUDIO CNN	0.98	0.1	0.80	0.3
HYBRID LSTM	0.80	0.4	0.78	0.5
HYBRID Bi-LSTM	0.88	0.2	0.76	0.2

Table 3
Evaluation parameters precision, recall, F1-score, support.

METHOD	PRECISION	RECALL	F1-SCORE	SUPPORT
TEXT CNN	0.63	0.68	0.60	33
AUDIO CNN	0.70	1.00	0.15	33
HYBRID LSTM	0.68	0.79	0.78	33
HYBRID-Bi-LSTM	0.75	0.73	0.74	33

Table 4
Parameters comparison with base paper [4].

METHOD	PRECISION	RECALL	F1-SCORE	SUPPORT
TEXT CNN	0.56	0.61	0.58	33
AUDIO CNN	0.47	0.3	0.13	33
HYBRID LSTM	0.48	0.65	0.55	33
Bi-LSTM	0.62	0.32	0.18	33

Table 5
Confusion matrix labels.

TRUE POSITIVE	TRUE NEGATIVE
FALSE POSITIVE	FALSE NEGATIVE

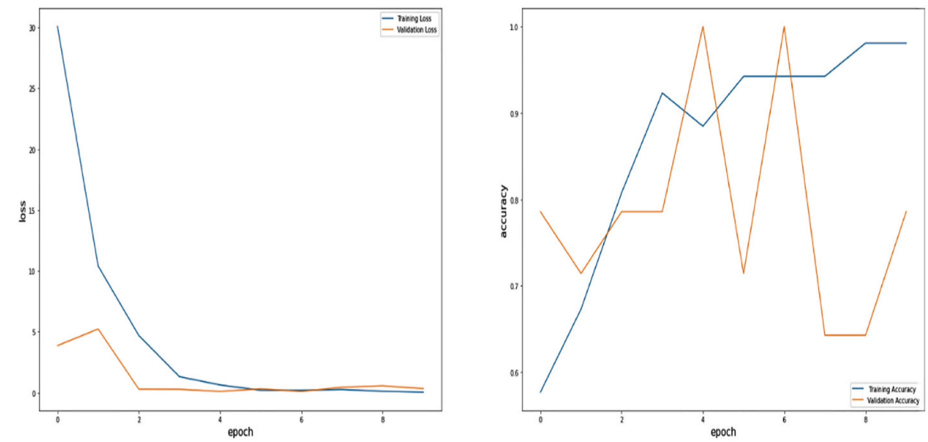


Fig. 5. Graphs of Text CNN model for Loss and Accuracy against no. of epochs. Fig. 5(a) represents Training Loss and Validation Loss against no. of epochs. Fig. 5(b) represents Training Accuracy and Validation Accuracy against no. of epochs.

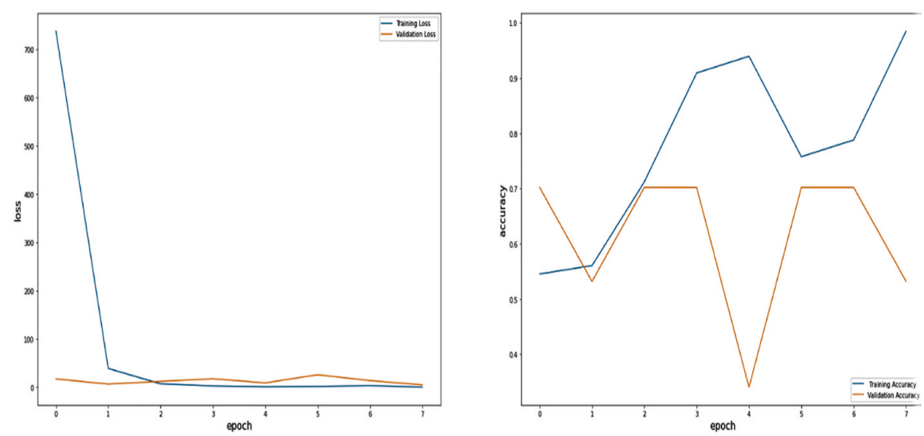


Fig. 6. Graphs of Audio CNN model for Loss and Accuracy against no. of epochs. Fig. 6(a) represents Training Loss and Validation Loss against no. of epochs. Fig. 6(b) represents Training Accuracy and Validation accuracy against no. of epochs.

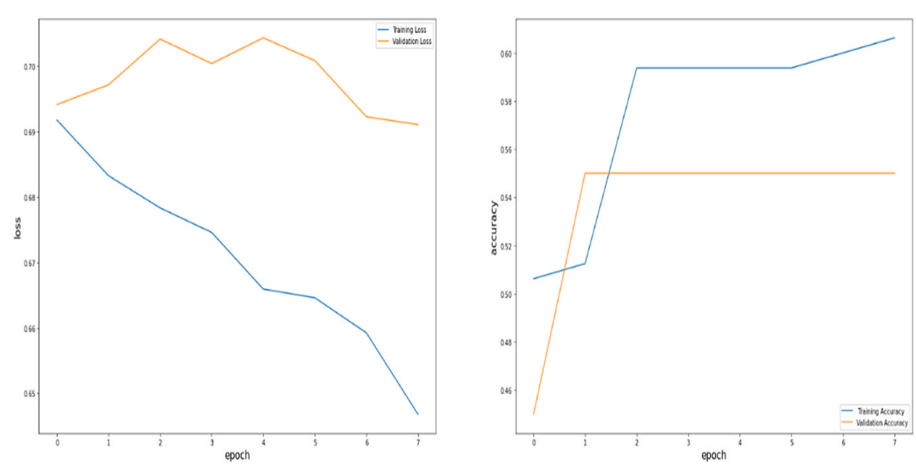


Fig. 7. Graphs of Hybrid LSTM model for Loss and accuracy against no. of epochs. Fig. 7(a) represents Training Loss and Validation Loss against no. of epochs. Fig. 7 (b) represents Training Accuracy and Validation Accuracy against no. of epochs.

loss and no. of epochs. Training loss and Validation loss is plotted. As number of epochs increases, training loss, decreases. In Figs. 5(b), 6(b) and 7(b), Fig 8(b) a graph is plotted between accuracy and no. of epochs. training accuracy and validation accuracy is plotted. Total no. of epochs is 10 in both cases. As epochs increases, Training accuracy also increases. It is observed from plotted graphs that audio CNN model has better learning rate than textual CNN that means audio CNN model has fast learning ability than textual CNN model. Loss of audio CNN model is very less as compared to textual CNN. Accuracy of audio CNN model has high learning features parameters therefore; it quickly reaches its peak value and then saturates very easily. In results, audio CNN model learns features faster as compared to text CNN model. In LSTM model accuracy

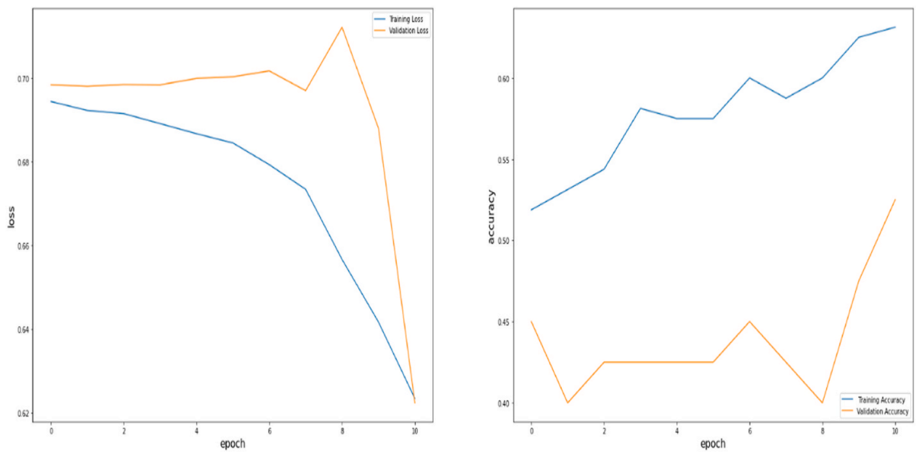


Fig. 8. Graphs of Hybrid Bi-LSTM model for Loss and accuracy against no. of epochs. Fig. 8(a) represents Training Loss and Validation Loss against no. of epochs. Fig. 8(b) represents Training Accuracy and Validation Accuracy against no. of epochs.

is less as compared to Bi-LSTM model. From graphs, it is clearly shows that Audio CNN model can predict depression with highest accuracy of 98% with minimum 1% loss. This means Audio CNN model can predict Depression easily; audio features can be detected by CNN model effectively. Also, Bi-LSTM model has high learning rate as compared to other models. Although, pre-processing time required for Bi-LSTM model is more as compared to other models, but with increase of no. of epochs accuracy also increases and loss decreases with no. of epochs. Therefore, from experimental results it is shown that, Bi-LSTM has higher learning rate and audio CNN predict accurately. Both the models are good according to their respective algorithm used.

4. Conclusion

In this work, a solution is proposed for an automatic depression detection system using deep learning. Three models are designed in the proposed work, first is textual CNN model, second is audio CNN model, third is hybrid LSTM and Bi-LSTM model. Hybrid model is a combination of audio and text modalities; therefore, it is named as a hybrid structure. In experimental results, it is shown that, for depression detection using deep learning, audio CNN model gives more accurate results in comparison to text CNN model, it can easily predict early symptoms of depression with an accuracy of 98% and loss of 0.1%, whereas text CNN give accuracy of 92% and loss of 0.2%. In confusion matrix, out of 47 people, text CNN predicts 20 as depressed and 3 as not-depressed whereas, audio CNN predict 22 as depressed and 5 as not-depressed. Similarly, from confusion matrix of LSTM model it is shown that, 26 are depressed patients and 6 are not-depressed whereas, from confusion matrix of Bi-LSTM model, 30 patients are depressed and only 5 as not depressed. Hence, it is proven that Bi-LSTM has better learning rates than LSTM model, which makes the model smart enough for remembering audio and text features. Confusion matrix of Bi-LSTM proves that deep learning has a solution for depression detection in patients. Also, training accuracy and validation accuracy of Bi-LSTM model is higher as compared to LSTM model. But, Bi-LSTM has disadvantage of more pre-processing time which makes the model good at learning but a slow model. Training time is more in Bi-LSTM model, whereas, LSTM is fast as compared to Bi-LSTM but a bad model for remembering audio and text features for a long period of time. LSTM model lost the past data whenever new information add in the model. Loss is minimum in all models which make accuracy of models high. In conclusion, prediction of depression is done with the help of graphs of training and validation and confusion matrix. Some parameters are also evaluated such as precision, recall, F1-score, support. More value of precision makes the model smart enough for future prediction and remembering past information and Bi-LSTM has maximum precision value. From Tables 6–9 it is clearly shown that, Bi-LSTM model and audio CNN has better values as compared to LSTM model and textual CNN model. At last, comparison of evaluation parameters of proposed model with base paper is done which indicates that, with same no. of epochs and same no. of layers, proposed model of audio CNN has better learning abilities than the model used in base paper [5].

5. Future scope

An automatic system for depression detection is proposed which predicts whether a person is depressed or normal person. For this purpose, a model is designed using CNN and LSTM algorithms. Some of the audio features and text features are extracted and applied in these algorithms. A hybrid model is designed using combination of audio and text features. Based on the objectives of proposed work, three models are designed i.e., audio CNN, textual CNN, hybrid LSTM and Bi-LSTM model. The proposed models have better learning rate and the hybrid model can be used for depression detection system. In the future, audio features, text features and also video features can be extracted and used for depression detection. Features can be extracted on better level in the future work. A hybrid model which combines audio, video and text features will be designed in future work. This will increase accuracy of proposed model and give better results than the previous models. More audio features are needed for accurate results. Other powerful algorithms will be applied on hybrid model like SVM, RF and Naïve Bayes. With more usage of powerful algorithms, make learning rate of proposed model better and hence more accurate results will be obtained by proposed model. For practical application, an application will be developed in future which will predict whether a person is depressed or normal person and predict severity of depression disease. This application will help users to do self-assessment and detect symptoms of depression automatically. Also, in future, more accurate and suitable dataset will be used for depression detection system. Compare results of various datasets will be occurred in future. Comparison of various algorithms such as CNN, LSTM, Bi-LSTM, SVM, RF and other machine learning and deep learning algorithms will be done for each dataset in the future. The proposed work will expand to real-time Depression detections system for people who are suffering from depression. Work will expand to identify depressed patients in social media platforms like Twitter and Reddit. In

Table 6
Confusion matrix of text CNN model.

Predicted Label/True Label	Depressed	Not-Depressed
Depressed	20	13
Not-Depressed	11	3

Table 7
Confusion matrix of audio CNN model.

Predicted Label/True Label	Depressed	Not-Depressed
Depressed	22	13
Not-Depressed	7	5

Table 8
Confusion matrix of hybrid LSTM model.

Predicted Label/True Label	Depressed	Not-Depressed
Depressed	26	7
Not-Depressed	8	6

Table 9
Confusion matrix of hybrid Bi- LSTM model.

Predicted Label/True Label	Depressed	Not-Depressed
Depressed	30	5
Not-Depressed	7	5

this, classification of posts based on stories post by user will be done. Classification is in the form of Binary labels: 0 (Depressed) and 1 (Not-Depressed).

CRediT authorship contribution statement

Vandana: Formal analysis, Investigation, Data curation, Writing – review & editing, Supervision, Project administration, All authors have read and agreed to the published version of the manuscript. **Nikhil Marriwala:** Conceptualization, Methodology, Software, Validation, Resources, Writing – original draft, preparation, All authors have read and agreed to the published version of the manuscript. **Deepti Chaudhary:** Visualization, Supervision, Project administration, All authors have read and agreed to the published version of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Depression - World Health Organization (WHO).".
[2] H.-N. Le, R.C. Boyd, Prevention of Major Depression: Early Detection and Early Intervention in the General Population, 2006, p. 23, 2022.
[3] Z. Wang, L. Chen, L. Wang, G. Diao, Recognition of audio depression based on convolutional neural network and generative antagonism network model, IEEE Access 8 (2020) 101181–101191, https://doi.org/10.1109/ACCESS.2020.2998532.
[4] N.S. Alghamdi, H.A. Hosni Mahmoud, A. Abraham, S.A. Alanazi, L. García-Hernández, Predicting depression symptoms in an Arabic psychological forum, IEEE Access 8 (2020) 57317–57334, https://doi.org/10.1109/ACCESS.2020.2981834.
[5] T. Alhanai, M. Ghassemi, J. Glass, Detecting depression with audio/text sequence modeling of interviews, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH (September) (2018) 1716–1720, https://doi.org/10.21437/Interspeech.2018-2522, 2018-Septe.
[6] L. Lin, X. Chen, Y. Shen, L. Zhang, Towards automatic depression detection: a bilstm/1d cnn-based model, Appl. Sci. 10 (23) (2020) 1–20, https://doi.org/10.3390/app10238701.
[7] N.V. Babu, E.G.M. Kanaga, Sentiment analysis in social media data for depression detection using artificial intelligence: a review, SN Comput. Sci. 3 (1) (2022) 1–20, https://doi.org/10.1007/s42979-021-00958-1.
[8] X. Ma, H. Yang, Q. Chen, D. Huang, Y. Wang, DepAudioNet: an efficient deep model for audio based depression classification, in: AVEC 2016 - Proc. 6th Int. Work. Audio/Visual Emot. Challenge, Co-located with ACM Multimed, 2016, pp. 35–42, https://doi.org/10.1145/2988257.2988267, 2016.
[9] B. Arroll, F. Goodyear-Smith, N. Kerse, T. Fishman, J. Gunn, Effect of the addition of a 'help' question to two screening questions on specificity for diagnosis of depression in general practice: diagnostic validity study, Br. Med. J. 331 (7521) (2005) 884–886, https://doi.org/10.1136/bmj.38607.464537.7C.
[10] J.F. Cohn, et al., Detecting depression from facial actions and vocal prosody, in: Proc. - 2009 3rd Int. Conf. Affect. Comput. Intell. Interact. Work. ACII 2009, 2009, https://doi.org/10.1109/ACII.2009.5349358.
[11] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, J. Krajewski, Automatic modelling of depressed speech: relevant features and relevance of gender, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 444 (2014) 1248–1252, https://doi.org/10.21437/interspeech.2014-313.
[12] D. Ramirez-Cifuentes, C. Largeron, J. Tissier, R. Baeza-Yates, A. Freire, Enhanced word embedding variations for the detection of substance abuse and mental health issues on social media writings, IEEE Access 9 (2021) 130449–130471, https://doi.org/10.1109/ACCESS.2021.3112102.
[13] G. Lam, H. Dongyan, W. Lin, S. City, Context; Multi-Modal, 2019, pp. 3946–3950.
[14] G. Rao, Y. Zhang, L. Zhang, Q. Cong, Z. Feng, MGL-CNN, A hierarchical posts representations model for identifying depressed individuals in online forums, IEEE Access 8 (2020) 32395–32403, https://doi.org/10.1109/ACCESS.2020.2973737.
[15] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T.F. Quatieri, A review of depression and suicide risk assessment using speech analysis, Speech Commun. 71 (June 2020) (2015) 10–49, https://doi.org/10.1016/j.specom.2015.03.004.
[16] R.L. Rosa, G.M. Schwartz, W.V. Ruggiero, D.Z. Rodriguez, A knowledge-based recommendation system that includes sentiment analysis and deep learning, IEEE Trans. Ind. Inf. 15 (4) (2019) 2124–2135, https://doi.org/10.1109/TII.2018.2867174.
[17] N. Cummins, V. Sethu, J. Epps, S. Schnieder, J. Krajewski, Analysis of acoustic space variability in speech affected by depression, Speech Commun. 75 (2015) 27–49, https://doi.org/10.1016/j.specom.2015.09.003.
[18] Y. Ding, X. Chen, Q. Fu, S. Zhong, A depression recognition method for college students using deep integrated support vector algorithm, IEEE Access 8 (2020) 75616–75629, https://doi.org/10.1109/ACCESS.2020.2987523.
[19] J.A. Sirey, et al., Perceived stigma as a predictor of treatment with depression, Am. J. Psychiatr. 158 (3) (2001) 479–481.
[20] I. Schumann, A. Schneider, C. Kantert, B. Löwe, K. Linde, Physicians' attitudes, diagnostic process and barriers regarding depression diagnosis in primary care: a systematic review of qualitative studies, Fam. Pract. 29 (3) (2012) 255–263, https://doi.org/10.1093/fampra/cmr092.
[21] J.R. Williamson, et al., Detecting depression using vocal, facial and semantic communication cues, in: AVEC 2016 - Proc. 6th Int. Work. Audio/Visual Emot. Challenge, Co-located with ACM Multimed, 2016, pp. 11–18, https://doi.org/10.1145/2988257.2988263, 2016.
[22] P. Corrigan, How stigma interferes with mental health care, Am. Psychol. 59 (7) (2004) 614–625, https://doi.org/10.1037/0003-066X.59.7.614.
[23] N. Marriwala, P. Rathee, An Approach to Increase the Wireless Sensor Network Lifetime, 2012, https://doi.org/10.1109/WICT.2012.6409128.
[24] L. Craft, J. Sport Exerc. Psychol. 20 (1998) 339–357. Craft-1998_Exercise-Clinical-Depression_Meta-Analysis_JSpExercPsychol.pdf.
[25] L. Yang, E. Pei, D. Jiang, M.C. Oveneke, X. Xia, H. Sahli, Multimodal measurement of depression using deep learning models, in: AVEC 2017 - Proc. 7th Annu. Work. Audio/Visual Emot. Challenge, 2017, pp. 53–54, https://doi.org/10.1145/3133944.3133948, *co-located with MM 2017*, no. January 2018.
[26] L. Yang, D. Jiang, H. Sahli, Integrating deep and shallow models for multi-modal depression analysis-hybrid architectures, IEEE Trans. Affect. Comput. 12 (1) (2021) 239–253, https://doi.org/10.1109/TAFFC.2018.2870398.
[27] L. Yang, D. Jiang, L. He, E. Pei, M.C. Oveneke, H. Sahli, Decision tree based depression classification from audio video and language information, in: AVEC 2016 - Proc. 6th Int. Work. Audio/Visual Emot. Challenge, Co-located with ACM Multimed. 2016, 2016, pp. 89–96, https://doi.org/10.1145/2988257.2988269. June 2018.
[28] Y. Yang, C. Fairbairn, J.F. Cohn, A. Member, From vocal prosody, IEEE Trans. Affect. Comput. 4 (2) (2013) 142–150.
[29] S. Yokoya, T. Maeno, N. Sakamoto, R. Goto, T. Maeno, A brief survey of public knowledge and stigma towards depression, J. Clin. Med. Res. 10 (3) (2018) 202–209, https://doi.org/10.14740/jocmr3282w.
[30] N.M. Vandana, Facial expression recognition using convolutional neural network, Mobile Radio Communications and 5G Networks, Lecture Notes in Networks and Systems 339 (2022) 605–617.
[31] G. Mendels, S.I. Levitan, K.Z. Lee, J. Hirschberg, Hybrid acoustic-lexical deep learning approach for deception detection, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH (2017) 1472–1476, https://doi.org/10.21437/Interspeech.2017-1723, 2017-Augus, no. November.
[32] H. Meng, D. Huang, H. Wang, H. Yang, M. Al-Shuraifi, Y. Wang, Depression recognition based on dynamic facial and vocal expression features using partial least square regression, in: *AVEC 2013 - Proc.* 3rd ACM Int. Work. Audio/Visual Emot. Chall., 2013, pp. 21–29, https://doi.org/10.1145/2512530.2512532. October.
[33] J.C. Mundt, A.P. Vogel, D.E. Feltner, W.R. Lenderking, Vocal acoustic biomarkers of depression severity and treatment response, Biol. Psychiatr. 72 (7) (2012) 580–587, https://doi.org/10.1016/j.biopsych.2012.03.015.
[34] L.S.A. Low, N.C. Maddage, M. Lech, L.B. Sheeber, N.B. Allen, Detection of clinical depression in adolescents' speech during family interactions, IEEE Trans. Biomed.

Eng. 58 (3 PART 1) (2011) 574–586, https://doi.org/10.1109/TBME.2010.2091640.
[35] T.E. Oxman, S.D. Rosenberg, P.P. Schnurr, G.J. Tucker, Diagnostic classification through content analysis of patients' speech, Am. J. Psychiatr. 145 (4) (1988) 464–468, https://doi.org/10.1176/ajp.145.4.464.
[36] A.D. Forfot, G. Østby, Digital Forensic Readiness in IoT - A Risk Assessment Model, vol. 1382, 2021, https://doi.org/10.1007/978-3-030-71711-7_5.
[37] Y. Meng, W. Speier, M. Ong, C.W. Arnold, HCET: hierarchical clinical embedding with topic modeling on electronic health records for predicting future depression, IEEE J. Biomed. Heal. Informatics 25 (4) (2021) 1265–1272, https://doi.org/10.1109/JBHI.2020.3004072.
[38] M.R. Morales, S. Scherer, R. Levitan, OpenMM: an open-source multimodal feature extraction tool, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH (September) (2017) 3354–3358, https://doi.org/10.21437/Interspeech.2017-1382, 2017-Augus.
[39] R. Rana, et al., Automated screening for distress: a perspective for the future, Eur. J. Cancer Care 28 (4) (2019) 1–13, https://doi.org/10.1111/ecc.13033.
[40] Y. Gong, C. Poellabauer, Topic modeling based multi-modal depression detection, in: AVEC 2017 - Proc. 7th Annu. Work. Audio/Visual Emot. Challenge, *Co-located with MM 2017*, 2017, pp. 69–76, https://doi.org/10.1145/3133944.3133945.
[41] S. Tariq, et al., A novel Co-Training-Based approach for the classification of mental illnesses using social media posts, IEEE Access 7 (2019) 166165–166172, https://doi.org/10.1109/ACCESS.2019.2953087.
[42] A.C. Trevino, T.F. Quatieri, N. Malyska, Phonologically-based biomarkers for major depressive disorder, EURASIP J. Appl. Signal Process. 2011 (1) (2011) 1–18, https://doi.org/10.1186/1687-6180-2011-42.
[43] A. Haque, M. Guo, A.S. Miner, L. Fei-Fei, Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions, 2018, pp. 1–7.
[44] J. Zinken, K. Zinken, J.C. Wilson, L. Butler, T. Skinner, Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression, Psychiatr. Res. 179 (2) (2010) 181–186, https://doi.org/10.1016/j.psychres.2010.04.011.
[45] Z. Zhao, et al., Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders, IEEE J. Sel. Top. Signal Process. 14 (2) (2020) 423–434, https://doi.org/10.1109/JSTSP.2019.2955012.
[46] J. Joshi, et al., Multimodal assistive technologies for depression diagnosis and monitoring, J. Multimodal User Interfaces 7 (3) (2013) 217–228, https://doi.org/10.1007/s12193-013-0123-2.
[47] M. Trotzek, S. Koitka, C.M. Friedrich, Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences, IEEE Trans. Knowl. Data Eng. 32 (3) (2020) 588–601, https://doi.org/10.1109/TKDE.2018.2885515.
[48] D. Analysis, I. Corpus, W.- Oz, DAIC - WOZ Depression Database, 2014.
[49] S. Alghowinem, et al., Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors, IEEE Transactions on Affective Computing 9 (4) (1 Oct.-Dec. 2018) 478–490, https://doi.org/10.1109/TAFFC.2016.2634527.
[50] N.B. Allen, S.E. Hetrick, J.G. Simmons, I.B. Hickie, What's the evidence? Early Intervention in Youth Mental Health 187 (7) (2007) 15–17.
[51] B. Sun, et al., A random forest regression method with selected-text feature for depression assessment, AVEC 2017 - Proc. 7th Annu. Work. Audio/Visual Emot. Challenge, co-located with MM 2017 (2017) 61–68, https://doi.org/10.1145/3133944.3133951.
[52] M.M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of depression-related posts in reddit social media forum, IEEE Access 7 (c) (2019) 44883–44893, https://doi.org/10.1109/ACCESS.2019.2909180.