Chapter 2
Speech Recognition-Based Prediction for Mental Health and Depression: A Review

Priti Gaikwad and Mithra Venkatesan

Abstract A person with a mental disorder exhibits a significant disturbance in his or her behavior. Generally, mental disorders are associated with distress or impairment of normal functioning. Lack of adequate resources and facilities, as well as a lack of awareness of the symptoms of mental illness, prevent people from getting the help they need. The ability to assess depression through speech is a critical factor in improving the diagnosis and treatment of depression. The spoken language is said to provide access to the mind, and a wide range of speech capture and processing technologies can be used to analyze mental health. Speech processing is about recognizing spoken words. The automatic recognition and extraction of information from speech enables the determination of some physiological characteristics that make a speaker unique to identify their mental health status. In this paper, we describe how mental health-related problems can be predicted by speech processing. This paper identifies the gaps in the literature review that lead to the proposed methodology.

Keywords Mental health · Speech processing · Natural language processing

2.1 Introduction

Mental health refers to the state of being aware of one's abilities, coping with daily stresses, working, and making a positive contribution to society. According to the World Health Organization (WHO), mental health is the absence of mental disorders. Our capacity to think, feel, interact with one another, and carry out daily tasks depends on both the mental health of each individual and the state of our society as a whole.

P. Gaikwad (✉) · M. Venkatesan
Dr. D. Y. Patil Institute of Technology, Pimpri Pune, Maharashtra, India
e-mail: ppgaikwad.scoe@sinhgad.edu

M. Venkatesan
e-mail: mithra.v@dypvp.edu.in

P. K. Jha et al. (eds.), *Proceedings of Congress on Control, Robotics, and Mechatronics*,
Smart Innovation, Systems and Technologies 364,
https://doi.org/10.1007/978-981-99-5180-2_2

Due to promotion, protection, and restoration, mental health has become a central concern of communities and societies around the world [1].

In humans, speech production is a result of physiological processes that are naturally affected by physical stress. There are major changes in the fundamental frequency level, the speaking rate, the pause pattern, and the breathiness of speech. A speech is one of the most natural and common ways in which we communicate with each other on a daily basis, and it contains a profound array of information that goes far beyond the verbal message it conveys. Listening to the speaker's utterances can reveal the speaker's gender, age, dialectal background, emotional status, and personality. A speaker's physiological and health condition is part of the paralinguistic information in his or her speech. Through signal processing techniques and statistical modeling, it is possible to capture natural changes in the human body by analyzing the sound and linguistic content of speech signals. Speech sounds in patients with depression tend to have a lower pitch, in the form of acoustic signals [2]. It is possible to detect and quantify disorders, diseases, monotonous speech, lower sound intensity, and slower speech rates, as well as more hesitations, stutters, and whispers. Speech has several advantages: it is difficult to hide symptoms, it directs emotion and thought through its language content, and it is an inexpensive medium. Due to similar vocal anatomy, it may generalize across languages, which is particularly useful when natural language processing technology is not available for low-resource languages. Since most clinical interviews are already recorded, it is easy to obtain using Smartphones, tablets, and computers rather than more costly wearable or invasive neuroimaging methods.

970 million people worldwide, or 1 in 8, experience mental disorders, primarily anxiety and depression. Due to COVID-19 pandemic, the number of people experienced anxiety and depression. According to preliminary projections, the prevalence of anxiety and major depressive disorders will rise by 26% and 28%, respectively, in 2020 [3].

There could be a "mental health epidemic" in India, according to President Ram Nath Kovind, who noted that 10% of the country's 1.3 billion people suffer from mental illness. The WHO estimates that about 15% of the world face the issue related to mental health in India. A meta-analysis of community surveys found that 33 out of 1,000 people experience depression or anxiety [4].

Physical health conditions like cancer, diabetes, and chronic pain can have underlying, life-altering effects on mental health conditions like stress, depression, and anxiety. So, given the aforementioned issue, depression must be automatically detected.

Thinking, feeling, and behavior are all impacted by depression. Depression makes day-to-day living more challenging and interferes with relationships, work, and study. If a person feels down, sad, or miserable most of the time for longer than two weeks, has lost interest in or pleasure from most of their usual activities, and exhibits multiple symptoms from at least three of the categories listed below, they may be depressed [5]. It is important to remember that everyone occasionally experiences some of these symptoms, and they may not signify depression per se. Likewise, not everyone who is depressed will exhibit all of these symptoms.

2.2 Literature Review

Researchers have developed many new methods proposed for speech patterns that indicate mental health. By analyzing depression detection, it can be seen that to evaluate patient health from an electronic health record is to map speech signals to depression features. Researchers proposed several models for detecting depression which leads to mental illness which are discussed below.

2.2.1 Related Work

In this study, Nanath et al. [6] have shown how social media data can be used to predict the mental health characteristics of people using text features and natural language processing. According to Alghowinem et al. [7], speech patterns, eye movements, and head posture were each analyzed for statistical features. A support vector machine (SVM) was used in emotion classification tasks. The Reddit database was used by Rssola et al. [8], who were able to identify trends in the writing style, emotional expression, and online behavior of the users in question by visualizing and analyzing some probabilistic features. Sarkara et al. [9] used emotions.csv dataset from the Kaggle Web site and used different machine learning (ML) and deep learning (DL) methods, multi-layer perceptron, convolution neural network, recurrent neural network with long short-term memory, SVM, and linear regression which are used as classifiers, to solve real-world problem; among them the RNN model has the highest accuracy 97.50% in the training set and in the test set 96.50%. Liu et al. [10] the goal of the NetHealth study was to predict people's mental health status by using network methods and DMF (a method from RS). Smartphone data, data from wearable sensors (Fitbit), and people's trait data from surveys were collected.

Due to behavioral interference from interviewers and problems in matching audio transcripts, Dong et. al. [11] only considered depression detection from non-interaction databases like DAIC-WOZ. In the future, it may be possible to use interaction databases to confirm the generalizability of the model. According to research by Ye et al. [12], patients with mild and minor depression have a higher recognition error rate than average individuals and patients with major depression. Di Matteo et al. [13] developed an Android app that collects regular audio recordings of participants' surroundings and recognizes English words with automatic speech recognition. Amanat et al. [14] obtained a large imbalanced dataset of tweets from the Kaggle Web site, implemented the one-hot coding method and principal component analysis (PCA), LSTM, and RNN for further improvement, and proposed a hybrid recurrent neural network for a large database. In the case of the real-time datasets, better results are obtained than other classification algorithms, which is in agreement with Gupta et al. [15]. However, the accuracy of the proposed algorithm can be improved for real-time recorded files by recording the speech in a professional environment and making an appropriate selection of neurons and values for drop-out layers.

Rejaibi et al. [16] worked with the dataset DAIC-WOZ, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). An LSTM was chosen for high-level audio feature extraction, suggesting that textual features may accurately represent depression. Rutowski et al. [17] use two corpora of American English speech collected by Ellipsis Health. They work with a deep learning model classifier based on deep NLP and transfer learning showed excellent transferability across age, gender, and ethnicity. Schultebraucks et al. [18] suggested that to fully incorporate the vast clinical knowledge of clinicians, larger samples are required to confirm the findings and test for interactions between verbal and facial modalities. Another drawback is the dependence of the feature extraction process on pre-trained models. Although the author used standard techniques, it is pointed out that facial expression recognition is known to have shortcomings and is subject to bias. Aloshban et al. [19] conducted the experiments using 59 interviews recorded in three psychiatric centers in Italy. Feature extraction is done by a bidirectional long-term memory network (BLSTM). The primary drawback of the study is that the interviews were manually transcribed, particularly in depressed patients, the results of the Beck Depression Inventory-II (BDI-II) are unreliable. El Shazly et al. [20] studied 48 Egyptian EFL learners, there is no control group, the sample is small, and the data are descriptive. Although the sample size was small, the use of a mixed methods design provides a better understanding. According to Garoufis et al. [21], a speech analysis system that acknowledges (anomalous) pre- and relapse states in persons with psychotic disorders utilizing unsupervised learning with convolutional autoencoders. Daus et al. [22] use the Linguistic Inquiry and Word Count (LIWC) method to analyze verbal information that has been automatically translated in terms of the number of words of emotional categories. According to Sharma et al. [23], the diagnosis of mental illnesses is based on standardized interviews with a deterministic set of questions and scales. The machine learning model created using the imbalance dataset results in predictions that are biased toward the majority class; as a result, the model will consistently forecast that depression is absent, even when it is. There are no agreed-upon and accepted standards for biomarker scores in various nations and ethnic groups, so the XGBoost model created under this study cannot be adapted to other nations and racial groups. Machine learning should be used to accurately study the different types of depression. According to Wang et al. [24], research should focus on extracting additional speech signal features to describe them more accurately and cross-language learning to increase the reuse rate of models. Villatoro-Tello et al. [25] worked on the (DAIC-WOZ) Alzheimer's dementia dataset. Bag-of-words (BoW), (LIWC), and Third BERT techniques were used. He suggested that the LA method can be fused with raw waveform based CNN to increase the performance.

According to Araño [26], it is challenging to infer happiness from speech characteristics. Future research can therefore concentrate on developing new descriptors that accurately depict this emotional state. Mou et al. [27] used CNN and LSTM and suggested expanding the sample size and participant age range to boost the model's generalizability. Unsupervised learning can identify driver stress-related features from unlabeled data.

Current research in the automatic detection of depression has a number of limitations. First, some methods rely heavily on manually selected questions, which require the involvement of psychologists with relevant expertise. In addition, the interview must cover all predetermined questions; otherwise, the analysis may be flawed. The question of how to enhance detection performance without pre-programmed questions remains a challenge. In addition, due to ethical concerns, there are not many publicly available depression datasets.

2.2.2 Gaps Identified

Despite the fact that speech depression recognition (SDR) has advanced using current datasets, the following are the dataset problems from the literature survey which impede its further advancement.

- Database annotation objectivity: Data annotation serves as the foundation for future work, but the performance of the developed model will be impacted by the distribution of depression values because annotators' perceptions are not always accurate.
- Small in scope and unavailability: Due to ethical issues and the sensitivity of depression speech, most institutions were unable to obtain sufficient samples. AVEC2013, AVEC2014, DAIC-WOZ, and BD are the only public depression databases currently available, and they are not suitable for scientific research. It is critical to address ethical concerns in publishing datasets.
- Non-universality: At the moment, interactive clinical interviews, where questions are carefully crafted such that there is no noise or interference. As a result, these data cannot represent depressed patients' daily lives accurately. Additionally, the issue of linguistic and cross-cultural communication has not yet been considered.
- Model generalization: The models are challenging to generalize to other datasets or data from different languages because the majority of studies only use one or a few small datasets. It is also necessary to improve the model's reliability and predictive validity across corpora, societies, languages, and crowded environments.
- Types of depression disorder: For instance, compared to the most prevalent major depressive disorder, the pathogenesis and behavior of bipolar disorder are different. Few studies have been conducted on how speech signals can differentiate between these two.
- Multi-modality fusion mechanism: Since various modalities can successfully complement one another, future research trends in depression analysis cannot be avoided, including the combining of multiple modalities. However, the success of multimodal research depends on an effective and appropriate mechanism.

According to the identified gaps in the literature, the following objectives have been formulated in the paper.

2.2.3 Objectives

- To study existing literature where speech and language processing is applied for mental well-being
- To build, collect, and process datasets based on speech for the diagnostic model
- To propose a diagnostic model capable of finding mental illness based on speech processing.

2.3 Proposed Methodology

The goal is to develop a safe speech diagnostic model for people with depression. Speech depression datasets are typically recorded by in-person, telephone, or virtual interviewers as clinical clinicians talk with depressed patients. Other modalities, including information from the depression scale, facial expressions, physiological dynamics, etc., are also sometimes recorded during data collection for supplemental analysis.

The proposed methodology is depicted in Fig. 2.1. The different levels involved in the model are database, preprocessing of speech data, feature extraction from speech data, and validation of model and classification as depressed or normal.

2.3.1 Dataset

The following datasets are public records.

The audio-visual depression language corpus for AVEC2013 includes the AVEC2013 and AVEC2014 datasets. AVEC2014 is a set of AVEC2013 consisting of 300 German videos with shorter video clips than in AVEC2013. One component of the Distress Analysis Interview Corpus used for AVEC2016 and AVEC2017 is the Distress Analysis Interview Corpus—Wizard of Oz (DAIC-WOZ).

The traditional method of diagnosing depression uses clinical interviews to screen potential patients for depression. However, these assessments rely heavily on physician questions, patient verbal reports, actions reported by family or friends, and mental status tests such as the Beck Depression Inventory, the Hamilton Rating Scale for Depression, and the Scale for the Assessment of Negative Symptoms. The PHQ-9 is the Health Status Questionnaire for measuring depression scale in that The DSM-nine IV diagnostic criteria for MDD. The PHQ-9 can be applied as a screening tool, a diagnostic tool, and a tool for symptom assessment. It can be used to track changes in particular symptoms over time as well as the overall extent of a patient's depression. Based on the depression score: 0–4 none, 5–9 mild, 10–14 moderate, 15–19 moderately severe, and 20–27 severe. These are all based on subjective assessments, and since there are no reliable, quantitative measures, the results often vary depending on the situation.

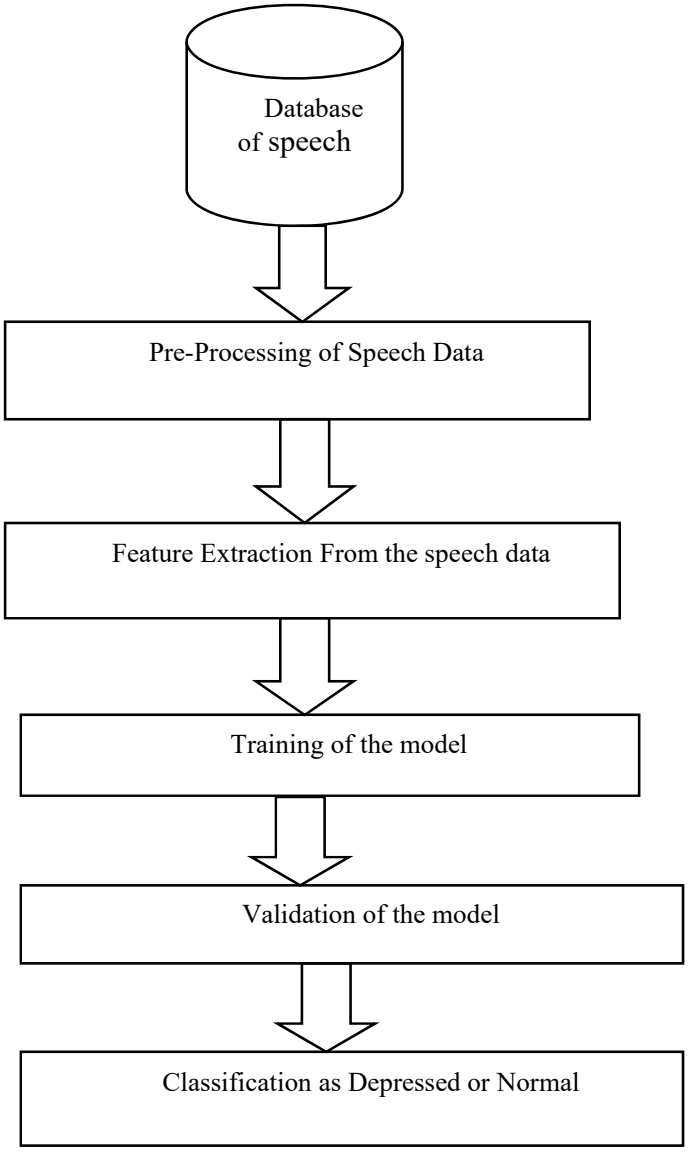

Fig. 2.1 Block diagram of prediction of mental health

The Distress Analysis Interview Corpus (DAIC) comprises an English database. It includes clinical interviews intended to assist in the identification of mental illnesses like anxiety, depression, or post-traumatic stress disorder. The "Wizard of Oz" interviews, conducted by a virtual interviewer, make up the depression section of the corpus. The PHQ-8 Depression Inventory, which is different from the databases

used in Germany and Turkey, was used to determine patients' depression scores. In total, there are 189 records of 189 patients.

Data Availability To download one of the databases users must complete the registration process and submit the form in order. Due to consent restrictions, only academics and other non-profit scholars are allowed to access datasets. Thus, when requesting, the user needs only provide their academic email address in order to download.

The DAIC-WOZ Database & Extended DAIC Database data available at https://dcapswoz.ict.usc.edu/

Extended DAIC Database The DAIC-WOZ database for the evaluation of PTSD and depression was created by ICT, and this is its expanded version. Additionally available upon request, this information was used for the AVEC 2019 Challenge.

We are going to use this E-DIAC English dataset for our proposed work, as it is secure and available as per request, it has large no of files, and annotation of the data is also done.

The E-DAIC is the next version of the DAIC-WOZ, collected from semi-clinical interviews to help with the treatment of mental disease like anxiety and depression. Participant information is labeled with age, gender, and PHQ-8 score, the dataset includes 163 developmental patterns, 56 samples for training, and 56 samples for testing. For AVEC2019, and this database will be used.

2.3.2 Preprocessing of Speech Data

Speech recognition faces many difficulties. First, there are not enough datasets in the field of speech, as the creation of a high-quality speech emotion database requires a lot of time and effort. Second, the different data in the database have different speakers, each with different gender, age, language, culture, and so on. Finally, sentences rather than specific words are often the basis for the emotions expressed in speech. Therefore, a challenge in current research is to increase the accuracy of emotion recognition by using low-level descriptors (LLDs) and sentence-level features. There are typically three methods in conventional techniques for speech emotion recognition. Data preprocessing, which includes data normalization, speech segmentation, and other operations, is the first step.

The original speech data must be enhanced by changing the speech playing speed, and the problem of an unbalanced distribution of speech data must be resolved using the balancing datasets weight method. We can use data enhancement and speech segmentation to increase the number of training samples to address the issue of the limited number of training samples.

2.3.3 Speech Feature Extraction

Features Speech features like Mel-frequency cepstral coefficients (MFCC), pitch, jitter, shimmer, energy, the zero-crossing rate (ZCR), the harmonic-to-noise ratio, fundamental frequency (F0), formant, low speech volume, monotone intonation, reduced articulation and the harmonic distribution, as well as perceptual linear prediction (PLP) coefficients have performed better in classifying an individual as a depressed or a healthy one.

The spectral features: related to spectral centroid; the cepstral features: related to the cepstrum analysis (an anagram to the spectrum signal) like the Mel-frequency cepstral coefficients (MFCCs); prosodic features: fundamental frequency F0 (the first signal harmonic) and the loudness the voice quality: like the formants (the spectrum maxima), the jitter (the signal fluctuation), and the shimmer (the peaks variation). Source feature: voice quality feature; deep audio feature: raw audio input for acoustic feature.

Techniques Recently, deep learning models have used convolution neural networks (CNNs) in particular in conjunction with automatic feature extraction, either explicitly from time-domain samples or using a frequency-domain representation of the signal, like the discrete Fourier transform (DFT) or spectrogram.

Successful approaches combine spectral features and their time derivatives with machine learning algorithms like hidden Markov models (HMMs), Gaussian mixture models (GMMs), or hybrid GMM-PLP coefficients. Convolution neural systems (CNNs) [4, 23] in particular have recently been used in deep learning models, utilizing either a frequency-domain representation of the signal or automatically extracting features from time-domain samples.

For spectral feature MFCC feature extraction MFCC-CNN, MFCC-RNN so we are proposed to use **MFCC Multichannel CNN-BLSTM** by fusion of magnitude and phase spectral feature.

We can combine spectral features with prosodic features, and an autocorrelation technique can extract pitch prediction like fundamental frequency (f0) raw. In prosodic features, we can extract probability of voicing (POV), F0 intensity, loudness, voice quality, and F0 envelope. The jitter and shimmer algorithm can be used for voice quality.

We can also use speech features like amplitude envelope, zero-crossing rate, and spectral flux.

2.3.4 Classification for Mental Health Data

To support the diagnosis of depression, it is therefore necessary to develop depression classification methods. Several techniques have recently been developed for assisting clinicians during the diagnosis and monitoring of clinical depression, the recent development of machine learning and artificial neural networks. Utilizing

sufficiently sizable speech corpora will allow for the development of the necessary models for mental health information prediction. Machine learning algorithms like SVM, decision trees, logistic regression, and KNN classifiers were used in the investigation. The model's accuracy was increased using the SMOTE method. On the other hand, deep learning models like CNN, ANN [4], and LSTM outperformed conventional machine learning techniques.

Deep learning models commonly used include CNN, LSTM, and BERT; however, BiLSTM [11] can offer higher accuracy. MT-CNN multitasks CNN [4] to perform better at predicting depression.

We can implement hierarchical depression models so that we can combine classification and regression for better performance parameters. In the hierarchical model in the first layer, multiple classifiers can be used as ensemble models and in the second layer for each recording regression algorithm can be used.

2.4 Discussion

The following points are suggested based on the thorough literature review and methodical meta-analysis. It details the approach taken, the benefits and difficulties encountered while using the datasets for depression. In comparison with volume-based features, SVM, multivariate regression, performs better depression prediction. A variety of mental illnesses can be quickly identified with the aid of RF, NB, and SVM. Spatiotemporal data can be extracted using the CNN and RNN combination. In order to detect depression more effectively and accurately, hybrid models like CNN with LSTM are used. Algorithms for machine learning and boosting are useful in identifying the sociodemographic and psychological factors that contribute to depression. It has been noted that the voice change study may aid in the early detection of depression.

2.5 Conclusion

Our lives depend on having a healthy mind. Serious issues brought on by mental instability are challenging to diagnose and treat. A serious mental health condition with high societal costs is depression. One of the objective indicators for the early detection of depression can be speech signal characteristics. To address the issue of the representation of speech signals by conventional feature extraction techniques is difficult; so in this study we proposed MFCC multichannel CNN-BLSTM spectral feature. We can add different speech features such as spectral feature, prosodic feature, and voice quality feature for extraction of the speech so that exact level of mental health can be identified. The potential of AI algorithms to address mental health questions in mental health care will give the best results.

References

1. Liu, S., Vahedian, F., Hachen, D., Lizardo, O., Poellabauer, C., Striegel, A., Milenković, T.: Heterogeneous network approach to predict individuals' mental health. ACM Trans. Knowl. Discov. Data **15**(2), Article 25
2. Stasak, B., Huang, Z., Joachim, D., Epps, J.: Automatic elicitation compliance for short-duration speech based depression detection. In: ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 978-1-7281-7605-5/20/$31.00 ©2021 IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414366
3. https://www.who.int/news-room/factsheets/detail/mental-disorders
4. https://www.ideasforindia.in/topics/human-development/understanding-india-s-mental-healthcrisis.html#:~:text=In%202017%2C%20the%20President%20of,49%20million%20from%20anxiety%20disorders
5. Priya, A., Garga, S., Tigga, N.P.: Predicting anxiety, depression and stress in modern life using machine learning algorithm. In: International Conference on Computational Intelligence and Data Science (ICCIDS 2019). Procedia Comput. Sci. **167**, 1258–1267 (2020)
6. Nanath, K., Balasubramanian, S., Shukla, V., Islam, N., Kaitheri, S.: Developing a mental health index using a machine learning approach: assessing the impact of mobility and lockdown during the COVID-19 pandemic. Technol. Forecast. Soc. Change **178**, 121560 (2022)
7. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Hyett, M., Parker, G., Breakspear, M.: Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. IEEE Trans. Affect. Comput. **9**(4) (2018)
8. Ríssola, E.A. Aliannejadi, M., Crestani, F.: Mental disorders on online social media through the lens of language and behaviour: analysis and visualization. Inf. Process. Manag. **59**, 102890 (2022)
9. Sarkara, A., Singh, A., Chakraborty, R.: A deep learning-based comparative study to track mental depression from EEG data. Neurosci. Inform. **2**, 772–5286 100039 (2022)
10. Liu, S., Vahedian, F., Hachen, D., Lizardo, O., Poellabauer, C., Striegel, A., Milenković, T.: Heterogeneous network approach to predict individuals' mental health. ACM Trans. Knowl. Discov. Data **15**(2), Article 25. Publication date: April 2021
11. Dong, Y., Yang, X.: A hierarchical depression detection model based on vocal and emotional cues. Neurocomputing **441**, 279–290 (2021)
12. Ye, J., Yu, Y., Wang, Q., Li, W., Liang, H., Zheng, Y., Fu, G.: Multi-modal depression detection based on emotional audio and evaluation text. J. Affect. Disord. **295**, 904–913 (2021)
13. Di Matteo, D., Fotinos, K., Lokuge, S., Mason, G., Sternat, T., Katzman, M.A., Rose, J.: Automated screening for social anxiety, generalized anxiety, and depression from objective smartphone-collected data: cross-sectional study. J. Med. Internet Res. **23**(8), e28918 (2021)
14. Amanat, A., Rizwan, M., Javed, A.R., Abdelhaq, M., Alsaqour, R., Pandya, S., Uddin, M.: Deep learning for depression detection from textual data. Electronics **11**, 676 (2022). https://doi.org/10.3390/electronics11050676
15. Gupta, M., Vaikole, S.: Audio signal based stress recognition system using AI and machine learning. J Algebraic Stat. 13(2), 1731–1740 (2022)
16. Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., Othmani, A.: MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. Biomed. Signal Process. Control **71**, 103107 (2022)
17. Rutowski, T., Shriberg, E., Harati, A., Lu, Y., Oliveira, R., Chlebek, P.: Cross-demographic portability of deep NLP-based. depression models. In: 2021 IEEE Spoken Language Technology Workshop (SLT), 978-1-7281-7066-4/20/$31.00 ©2021 IEEE. https://doi.org/10.1109/SLT48900.2021.9383609
18. Schultebraucks, K., Yadav, V., Shalev, A.Y. Bonanno, G.A., Galatzer-Levy, I.R.: Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. PsychologicalMedicine 1–11. https://doi.org/10.1017/S0033291720002718

19. Aloshban, N., Esposito, A., Vinciarelli, A., What you say or how you say it? Depression detection through joint modeling of linguistic and acoustic aspects of speech. Cognitive Comput. https://doi.org/10.1007/s12559-020-09808-3
20. El Shazly, R.: Effects of artificial intelligence on English speaking anxiety and speaking performance: a case study. Expert Syst. **38**, e12667 (2021)
21. Garoufis, C., Zlatintsi, A., Filntisis, P.P., Efthymiou, N., Kalisperakis, E., Garyfalli. V., Karantinos, T., Mantonakis, L., Smyrnis N., Maragos, P.: An unsupervised learning approach for detecting relapses from spontaneous speech in patients with psychosis. In: Proceedings 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, July 2021
22. Daus, H., Backenstrass, M. Feasibility and acceptability of a mobile-based emotion recognition approach for bipolar disorder. Int. J. Interact. Multim. Artif. Intell. **7**(2)
23. Sharma, A., Verbeke, W.J.M.I.: Improving diagnosis of depression with XGBOOST machine learning model and a large biomarkers Dutch Dataset ($n = 11{,}081$). Front. Big Data. **3**, Article 15 (2020). www.frontiersin.org
24. Wang, H., Liu, Y., Zhen, X., Tu. X.: Depression speech recognition with a three-dimensional convolutional network. Front. Hum. Neurosci. **15**, Article 713823 (2021). www.frontiersin.org
25. Villatoro-Tello, E., Pavankumar Dubagunta, S., Fritsch, J., Ramírez-de-la-Rosa, G., Motlicek, P., Magimai-Doss, M.: Late Fusion of the available lexicon and raw waveform-based acoustic modeling for depression and dementia recognition
26. Araño, K.A., Gloor,· P., Orsenigo, C., Vercellis, C.: When old meets new: emotion recognition from speech signals. Cognitive Comput. **13**, 771–783 (2021). https://doi.org/10.1007/s12559-021-09865-2
27. Mou, L., Zhou, C., Zhao, P., Nakisa, B., Rastgoo, M.N., Jain, R., Gao, W.: Driver stress detection via multimodal fusion using attention-based CNN-LSTM. Expert Syst. Appl. **173**, 114693 (2021)