A Comprehensive Analysis of Speech Depression Recognition Systems

Ali Hassan
Department of Electrical and Computer Engineering
Florida Agriculture and Mechanical University
Tallahassee, Florida, USA
ali1.hassan@famu.edu

Shonda Bernadin
Department of Electrical and Computer Engineering
Florida Agriculture and Mechanical University
Tallahassee, Florida, USA
bernadin@eng.famu.fsu.edu

***Abstract*—Being the third most common cause of disability globally, clinical depression is a serious global health concern that is characterized by melancholy, loneliness, and low self-esteem. About 10% of adults in the US alone suffer from this mental disorder, which is difficult to quantify because it is subjective. The subjectivity of traditional diagnostic techniques like surveys and interviews is a drawback. While more objective, biological markers run the risk of incorrect diagnosis. To highlight the distinctive acoustic characteristics of depressed people's speech, such as pauses, low energy, and monotonicity, this paper investigates the possibility of speech patterns serving as objective markers for depression. It talks about how research on Speech Depression Recognition (SDR) is moving toward deep learning models such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN). The difficulties encountered in SDR research are also discussed in the paper, such as the requirement for sizable, trustworthy datasets and the shortcomings of the available databases in terms of scenario diversity, imprecise labeling, and privacy restrictions. To conduct a more precise and effective analysis of depression, the conclusion highlights the significance of comprehending the physiological effects of depression on speech, improving data collection, fostering interdisciplinary collaboration, investigating various forms of depression, and integrating multimodal data.**

***Index Terms*—Clinical Depression, Speech Patterns, Speech Depression Recognition, Acoustic Features, Deep Learning, Convolutional Neural Networks, Long Short-Term Memory Networks, Diagnostic Methods, Mental Health.**

I. Introduction

Clinical depression, often characterized by sadness, loneliness, and low self-esteem (1), ranks as the third leading cause of disability and suicide globally. In the US, about 10% of adults suffer from this condition (2). Unlike straightforward physical health measures, depression's impact is complex and multifaceted.

WebMD's research highlights that untreated depression can lead to harmful behaviors like substance abuse and alcoholism, adversely affecting personal and professional relationships (3). Prolonged exposure to stressors can also trigger physiological responses, including increased cortisol levels and altered cardiac function, potentially leading to cardiovascular diseases and metabolic disorders (3). Moreover, depression can disrupt sleep patterns, causing further emotional and physical distress, including significant weight changes (3).

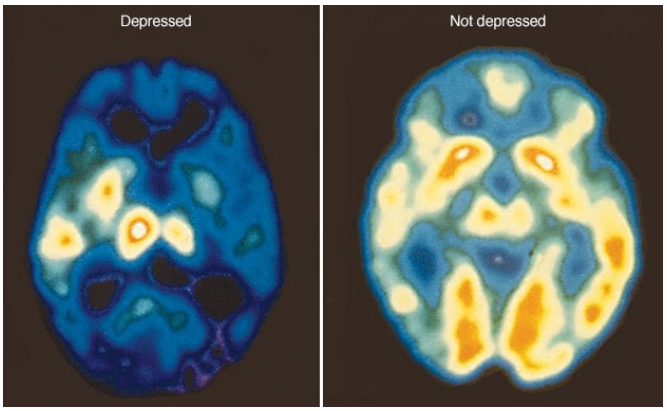

Fig. 1: *Diagnostic Comparison of Brain Scans: A Study of Depressive Symptomatology Using PET Imaging(5)*

A. Impacts of Depression

1) Depressed Brain: Depression can lead to permanent brain changes, impacting memory and concentration, with studies showing reduced size in specific brain regions in clinical depression cases, sometimes reversible (4). Key affected areas include the hippocampus (memory, learning), thalamus (information transfer), amygdala (emotion regulation), and prefrontal cortex (attention, emotions), with abnormal interac-

SoutheastCon 2024 | 979-8-3503-1710-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/SOUTHEASTCON52093.2024.10500078

tions increasing stress hormones and causing further damage (6; 7).

B. Impacts of Depression on Individuals Life

As mentioned in previous sections, the volume of different parts of the brain shrinks due to clinical depression. One study found that amygdala-prefrontal cortical dysfunction is the cause of the following symptoms shown in depressed people(8): Emotional regulation is lost such as feeling sad, hopeless, and tearful, Loss of interest in many enjoyable activities, Having slowed down physical actions as well as thoughts, Reduction of empathy in individuals, Indecisiveness and difficulty concentrating/thinking because of shrinkage of the hippocampus

In 2004, the World Health Organization predicted it to become 2nd leading cause of diseases worldwide by 2030 (9).

C. Impacts of Depression worldwide

Depression affects health and the economy, with Olesen et al. estimating the 2010 cost in Europe at €24,000 per patient (€92 billion total, €54.104 billion from lost productivity) (10), Stewart noting a $44 billion loss in the US (2002) due to reduced productivity (10), and the WHO reporting over 800,000 suicide deaths annually, affecting six people each, underlining depression's broad socio-economic impact (11)(12).

II. Current Diagnostic Methods

The National Institute of Mental Health (NIMH) highlights that Major Depressive Disorder often goes undiagnosed, with 48.4%(14) of depression cases unnoticed, necessitating the use of tools to detect symptoms such as sleep issues, lack of interest, guilt, energy loss, focus difficulties, appetite shifts, psychomotor alterations, and thoughts of suicide, summarized as SIGE CAPS. Essential diagnostic methods include the self-administered Personal Health Questionnaire (PHQ-8) for evaluating depression severity(15), the Beck Depression Inventory (BDI) with 21 questions to measure depression's depth(16), and the clinician-led Hamilton Depression Rating Scale (HAM-D) assessing depression and suicidality(17).

The section discusses *self-administered or interview-style tests* for depression, noting interview tests are influenced by the doctor's subjectivity (18), while self-administered tests are subject to the Hawthorne effect, where individuals alter their behavior when observed (19), suggesting the need for more objective depression detection methods.

TABLE I: *Table showing different depression tests showing scores for severity of depression*

Test	Normal	Slight	Medium	Extreme	Very Extreme
HAM-D	0-7	8-13	14-18	19-22	≥ 23
BDI-II	0-13	14-19	20-28	29-63	-
PHQ-8	0-4	5-9	10-14	15-19	20-24
PHQ-9	0-4	5-9	10-14	15-19	20-27
YMRS	0-5	6-12	13-19	20-29	≥ 30
MADRS	0-11	12-22	23-30	31-35	≥ 36

III. Speech as an Objective marker

Cognitively, the production of speech involves the following steps as shown in Figure 2:

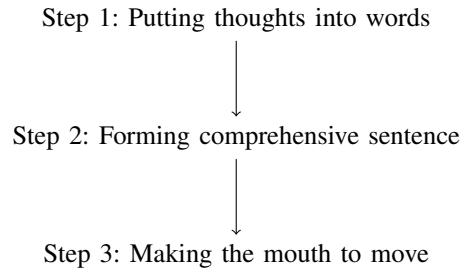

Fig. 2: *The process of formation of speech*

Human speech processing involves specific brain areas, including Broca's area(22) in the left frontal lobe for converting thoughts into words, and the motor cortex for mouth movements, with impairments affecting communication. Depression alters speech characteristics, making it lower, monotonic, and labored, with increased pauses(23), absolutes, self-focused, and negative language, as per the Journal of Neurolinguistics and Clinical Psychological Science. Recent studies reveal a distinct acoustic profile in depressed speech, with changes in fundamental frequency and formants, highlighting acoustic features as potential objective markers for depression detection (24)(25)(26).

IV. Depression Detection Using Audio

Audio data, as a non-invasive, clinically accessible detector of depression, uses objective acoustic features unlike subjective measures, combining standardized (questionnaires, EEG) and non-standardized (behavior, expressions, vocal inflection) data in speech signal analysis to assess depression presence and severity(28).

The labeled speech signal sample is represented as follows:

$$[signal_i, label_i], i \in N \tag{1}$$

TABLE II: *Comparative Analysis of Vocal Attributes Across Different Emotional States: Anger, Happiness, Sadness, Fear, and Disgust* (27)

	Anger	**Happiness**	**Sadness**	**Fear**	**Disgust**
Speech rate	Slightly faster	Faster or slower	Slightly Slower	Much Slower	Very Much Slower
Pitch Average	Very much higher	Much Higher	Slightly Slower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy chest tone	Breathy, blaring	Resonant	Irregular voicing	Grumbled chest tone
Pitch Change	Abrupt, on stressed syllables	Smooth, upward inflections	Downward inflections	Normal	Wide downward terminal inflections
Articulation	Tense	Normal	Slurring	Precise	Normal

Critical to this approach is evaluating the relationship between the mapping functions f and g, ensuring correlation between speech signals and depression features as shown in:

$$f(g(signal_i)) \rightarrow label_i \tag{2}$$

This process involves analyzing speech signals $signal_i$ and their depression-related characteristics using the function g, with a function f mapping these features to corresponding labels (Equation 1 and 2). Traditional features like fundamental frequency have limitations in Speech Depression Recognition (SDR) systems (29). Hence, deep learning techniques are employed for extracting sophisticated features, leading to enhanced detection capabilities as detailed in Figure 3.

V. Speech Depression Classification Methods

Previous studies on the SDR system utilized conventional algorithms like SVM, GMM, or LR post feature extraction, reflecting the broader use of AI in fields such as additive manufacturing (95).

In Figure 5, a developmental history of SDR can be seen.

AVEC 2016 researchers utilized a Support Vector Machine (SVM) with audio features for depression classification, optimizing it with Grid search for AUC and F1 scores (55). Extending Gong et al.'s work (56), they added context-specific features across modalities using a linear SVM. Zaremba et al. introduced a logistic regression outperforming GMM and SVMs in accuracy (57). Previous studies used GMM (58)(59)(60), while Gaussian Staircase Regression (GSR) and ensemble Gaussian classifiers later enhanced prediction over traditional methods, linking depression scores with speech features (61)(62)(63).

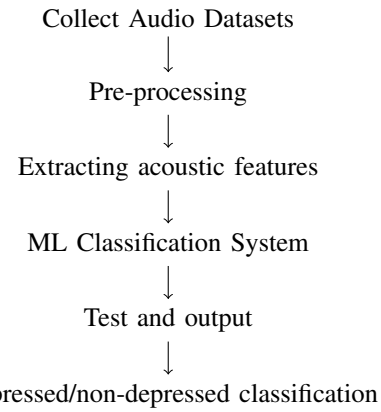

Fig. 3: *Process of Depression Detection using Audio*

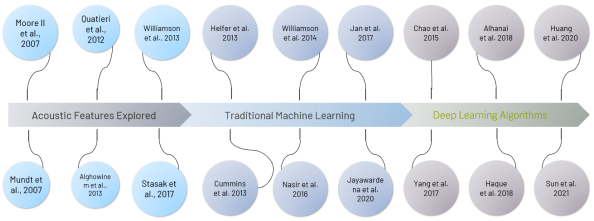

Fig. 4: *Developmental history of SDRs*

TABLE III: *Summary of traditional machine learning-based classification algorithms.*

Method	**Papers**	**Dataset**	**Performance**
SVM	Yuan et al. (45)	DAIZ-WOZ	MAE/RMSE 3.96/4.99
	Meng et al. (46)	DAIC-WOZ	F1 0.63
	Dan et al. (47)	DAIC-WOZ	-
	Cummins et al. (48)	Mundt-35	Accuracy 66.9%
GMM	JR Williamson et al. (49)	AVEC2014	MAE/RMSE 6.52/8.50
	Williamson et al. (50)	AVEC2013	MAE/RMSE 5.75/7.42
	Mehta et al. (51)	Mundt-35	AUC 0.76
Decision Tree	Quatieri et al. (52)	DAIZ-WOZ	F1 (D/N)0.52/0.81
LR	E.Harrell et al. (53)	DAIC-WOZ	RMSE 6.84
	Jan et al. (54)	AVEC2014	MAE/RMSE 6.14/7.43

VI. Datasets for SDR System

Data is essential for developing, training, and testing machine learning algorithms, with better data enhancing performance; SDR systems use multimodal datasets (audio, video, text) for depression detection (see Table V), mainly from interviews (Table IV) between doctors and patients via phone, virtual, or direct conversations.

TABLE IV: *Types of interviews for audio/video recordings*

Type	Description
Face-to-face	Participants answer predefined questions in person.
Teleconference	Phone interviews by human interviewers.
Wizard-of-Oz	Human-controlled agent, Ellie, conducts face-to-face interviews.
Automated	Autonomous agent, Ellie, conducts face-to-face interviews.

The datasets utilized in this study are comprised of three essential components:

1) Patient-centered interactive interviews
2) Pictorial representation and their description
3) Written or spoken text

Studies show gender-specific SDR systems benefit from divided datasets for depression detection, with males excelling in picture descriptions and females in interactive interviews, highlighting the need for gender consideration in SDR system design.

TABLE V: *Speech depression recognition system dataset resources* (30)

Data Corpus	Manner	Sample Size	Amount of clips	Timeframe
MODMA (2020) (31)	Audio/EEG	29 non-depressed, 23 depressed	1508 clips	Max 2.45 min
Bipolar corpus (2018) (32)	Audio/Video	46 non-depressed, 46 depressed	218 clips	Max 3.7 min
E-DAIC (2014) (33)	Audio/Video	351 depressed	275 clips	-
DAIC-WOZ (2014) (34)	Audio/Video/ECG/GSR	189 depressed	189 clips	5–20m, 15–25m
AVEC2014 (35)	Audio/Video	84 depressed	300 clips	6s-4 m
AVEC2013 (36)	Audio/Video	84 depressed	150 clips	20–50m
Mundt-35 (2007) (37)	Audio	35 depressed	-	-

The Audio-Visual Depression Language Corpus (AVDLC), combining AVEC2013 and AVEC2014, is key for emotion analysis in multimedia, with AVEC2013's 340 German videos and AVEC2014's 300 clips annotated for depression levels (38). The Distress Analysis Interview Corpus (DAIC), notably DAIC-WOZ for AVEC2016 and AVEC2017, merges audio, video, and physiological data for mental health studies, enhanced by E-DAIC's demographic and PHQ-8 scores (39). The Turkish Audio-Video Bipolar Disorder Corpus aids AVEC2018's bipolar research with 30 annotated videos, while the MODMA dataset introduces a Chinese multi-modality corpus with EEG and audio for psychiatric research.

A. Data Distribution

The DAIC-WOZ dataset, crucial for analysis, splits into training (107 audio files; 21 depressed, 86 non-depressed), development (35; 7 depressed, 28 non-depressed), and testing (47) sets, excluding an evaluation set; PHQ-8 scores depicted in Figures 5 (training, left) and (development, right).

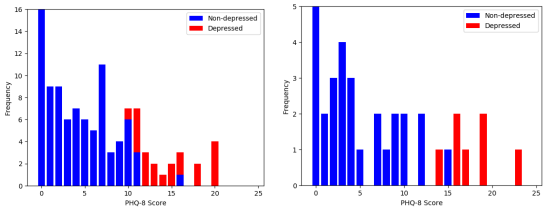

Fig. 5: *PHQ-8 score distribution in the train set (left) and development set (right).*

Figure 6 (left) depicts the depression level distribution in the train set, while Figure 6 (right) depicts the distribution in the development set (right).

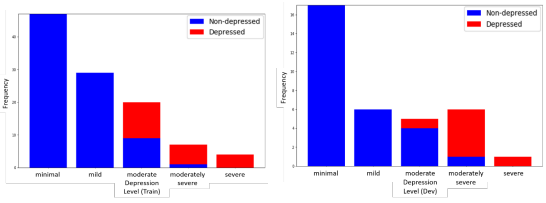

Fig. 6: *Depression level distribution in Train Set (Left); Depression level distribution in Dev Set (Right)*

VII. Hand-crafted Audio Features

Depressed speech characteristics—such as lower, monotonic, and labored speech with pauses and negative language—are detectable via deep learning through audio features like signal energy, entropy, and zero-crossing rate,

indicating speech loudness variation and flow. These features are essential for machine learning in intelligent systems, categorized by their level of abstraction, temporal scope, musical aspect, signal domain, and machine learning approach, including both hand-crafted and automatically extracted features(40). Key acoustic features for speech analysis include prosodic features (tracking pitch and loudness variations), voice quality features (measuring lung airflow and vocal tract motion), formant features (analyzing vocal tract resonance), and spectral features (reflecting vocal tract articulator movements)(41; 42; 43; 44).

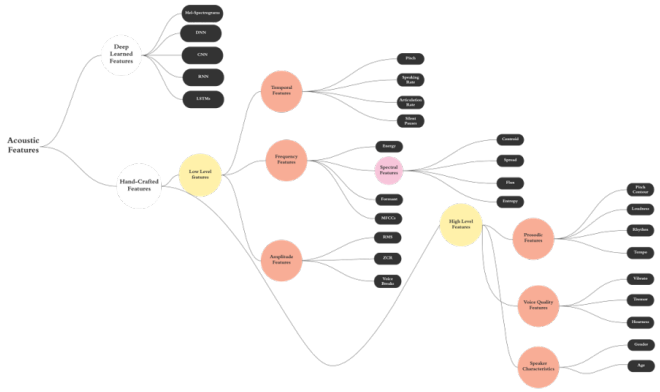

Fig. 7: *Flowchart of Acoustic Feature Analysis for Speech Depression Recognition.*

VIII. Speech Depression Recognition using Deep Learning

Deep learning, especially neural networks, has propelled speech-emotion recognition forward. Deep Neural Networks (DNNs), as Meyer et al. (64) have shown, and the DNN-ELM method by Han et al. (65), outperform traditional methods by extracting high-level features from speech. Despite their success, DNNs' dependency on vocal expressions and context poses real-world challenges, partially mitigated by Bertero et al. through Convolutional Neural Networks (CNNs) (66), which excel in emotion recognition despite simplicity. Recurrent Neural Networks (RNNs), researched by Park et al. (67), excel in handling temporal data, with LSTM, GRU, and QRNN improving efficiency and mitigating overfitting, essential for small emotional datasets. The integration of CNN and RNN into CRNN frameworks marks a recent advancement, aiming for precise emotion detection by focusing on low-dimensional features.

A. Research using Deep Learned features

Pre-trained network features, especially ResNet's for speech-to-text, show noise resilience (68). FVTC-CNN and DCNN-DNN models predict depression by analyzing neural network patterns (69; 70). Dual-layer networks and auto-encoders, including variants like de-noising and convolutional, enhance feature extraction and temporal analysis, crucial for unsupervised learning (71; 72; 73).

B. Deep Classifiers

Various deep classifier algorithms such as Recurrent Neural Networks (RNN), Deep Belief Networks, and Convolutional Neural Networks (CNN) are employed in Speech Depression Recognition, as depicted in Table VI.

TABLE VI: *Deep Classifiers applied in SDR*

Method	Papers	Dataset	Performance
RNN	Chao et al. (74)	AVEC2014	MAE/RMSE 7.91/9.98
	Al Jazraey et al.(75)	AVEC2013/14	MAE/RMSE 7.37/9.28
CNN	Yang et al. (76)	DAIC-WOZ	MAE/RMSE 5.163/5.974
	Haque et al.(77)	DAIC-WOZ	F1/Precision/Recall 0.769/71.4%/83.3%
	Lang He et al.(78)	AVEC2013/14	MAE/RMSE 8.78/10.90
LSTM	Alhanai et al. (79)	DAIC-WOZ	MAE/RMSE 4.97/6.27
	Zhangyin et al.(80)	BD	UAR/UAP/Accuracy 0.651/0.678/65.0%
	Salekin et al.(81)	DAIC-WOZ	F1/Accuracy 0.901/90%

CNNs excel in speech classification by capturing acoustic features for depression indicators(82; 83; 84), unlike LSTMs struggling with long sequences; Hague et al.'s(85) C-CNN enhances SDR through a multi-modality approach, outperforming traditional CNNs(86), while Niu et al.'s(87) TFCA block targets depression's time-frequency features, and RNNs surpass LSTMs in SDR despite deep network challenges (88; 89; 90).

IX. End to End Deep Architectures

End-to-end deep learning models in speech depression recognition (SDR) directly input raw data for training, improving feature learning and classification without prior knowledge. Despite their benefits, these models often struggle with large datasets and can be prone to overfitting and lack clarity in their decision-making processes.

Ma et al. devised DepAudioNet, a deep learning model integrating LSTM and DCNN, surpassing traditional acoustic feature-based SDR approaches (91). The model harnesses CNN for high-level feature extraction and LSTM for analyzing temporal aspects of audio signals. Tested on the DAIC-WOZ dataset, this innovative approach effectively addresses the instability of Mel scale features and enhances end-to-end SDR models. A randomized sample selection strategy was employed to counteract pattern distribution disparities.

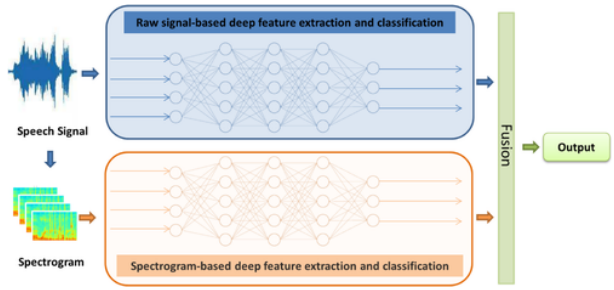

Fig. 8: *Deep architecture framework for end-to-end processing* (91)

TABLE VII: *Performance of Mel-scale filter bank features with different parameters for DepAudioNet (values of non-depression in brackets) (92)*

Partition	Time Window W	Max-pooling length l	F1 score	Precision	Recall
Baseline(Valstaret al., 2016)		-	0.41(0.58)	0.27(0.94)	0.89(0.42)
Development	120	3	0.52(0.70)	0.35(1.00)	1.00(0.54)

X. Challenges & Issues in SDR Research

A. Key Challenges

This section outlines key challenges in SDR research and future study implications. It emphasizes the necessity of accessible, high-quality data sets and patient cooperation, underlined by data privacy laws like HIPAA and personal freedom principles. Ethical adherence, highlighted by the Helsinki Declaration, and multi-level patient consent are critical. The promotion of open, standardized, and machine-readable research data encounters organizational and regulatory hurdles, with ethical and legal compliance being paramount. Additionally, enhancing SDR model generalizability across diverse data sets and ensuring annotation accuracy through rigorous protocols are fundamental for achieving reliable outcomes.

B. Prevailing Issues

1) Annotation Objectivity, Data Limitations, and Generalizability Constraints: Subjectivity in depression and speech variability challenges annotation objectivity, necessitating strict protocols for inter-annotator agreement. Ethical concerns and the lack of open databases restrict dataset sizes, impacting SDR development. Furthermore, clinical interview datasets may not accurately reflect real-life depression symptoms, and the absence of cross-cultural considerations limits SDR model generalizability across different populations.

XI. Conclusion

Clinical depression, a major disability caused globally, demands precise diagnosis with current methods (surveys, interviews) being subjective. Biological markers provide an objective but risky alternative. Research highlights depression's speech patterns (e.g., pauses, low energy, monotonicity) with acoustic features like energy and spectral attributes as objective indicators.

Advancements in Speech Depression Recognition (SDR) have shifted from hand-crafted features to deep learning (CNN, LSTM), requiring large, diverse, and accurate datasets. However, existing databases face issues such as diversity, labeling accuracy, and privacy, highlighting the need for a comprehensive, public database.

Future SDR research should aim at understanding depression's impact on speech, improving data collection, and fostering interdisciplinary efforts. It should also explore various depression types and utilize multimodal data for better analysis accuracy, emphasizing the importance of efficient multimodal integration.

References

[1] Getsmartacre, "What is the history of depression?," BrainsWay, 09-Jun-2022. [Online]. Available: https://www.brainsway.com/knowledge-center/the-history-of-depression/. [Accessed: 08-Oct-2022].

[2] "Mental illness," National Institute of Mental Health. [Online]. Available: https://www.nimh.nih.gov/health/statistics/mental-illness. [Accessed: 08-Oct-2022].

[3] M. D. Daniel K. Hall-Flavin, "Severe, persistent depression," Mayo Clinic, 13-May-2017. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/depression/expert-answers/clinical-depression/faq-20057770. [Accessed: 08-Oct-2022].

[4] P. Gorwood, "Neurobiological mechanisms of anhedonia," Dialogues in Clinical Neuroscience, vol. 10, no. 3, pp. 291–299, 2008.

[5] M. Clinic, PET scan of the brain for depression. .

[6] E. Palazidou, "The neurobiology of depression," British Medical Bulletin, vol. 101, no. 1, pp. 127–145, 2012.

[7] "What is depression? - Helen M. Farrell," TED. [Online]. Available: https://ed.ted.com/lessons/what-is-depression-helen-m-farrell. [Accessed: 08-Oct-2022].

[8] E. L. Moses-Kolko, S. B. Perlman, K. L. Wisner, J. James, A. T. Saul, and M. L. Phillips, "Abnormally reduced dorsomedial prefrontal cortical activity and effective connectivity with amygdala in response to

negative emotional faces in postpartum depression," American Journal of Psychiatry, vol. 167, no. 11, pp. 1373–1380, 2010.

[9] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," PLoS Medicine, vol. 3, no. 11, 2006.

[10] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T. F. (2015). A Review of Depression and Suicide Risk Assessment Using Speech Analysis. Speech Communication, 71, 10–49. http://doi.org/10.1016/j.specom.2015.03.004

[11] W. F. Stewart, "Cost of lost productive work time among US workers with depression," JAMA, vol. 289, no. 23, p. 3135, 2003.

[12] A.-F. Näher, C. Rummel-Kluge, and U. Hegerl, "Associations of suicide rates with socioeconomic status and social isolation: Findings from longitudinal register and Census Data," Frontiers in Psychiatry, vol. 10, 2020.

[13] "Depression," National Institute of Mental Health. [Online]. Available: https://www.nimh.nih.gov/health/topics/depression. [Accessed: 08-Oct-2022].

[14] S. Z. Williams, G. S. Chung, and P. A. Muennig, "Undiagnosed depression: A community diagnosis," SSM - Population Health, vol. 3, pp. 633–638, 2017.

[15] S. S. Dhingra, K. Kroenke, M. M. Zack, T. W. Strine, and L. S. Balluz, "PHQ-8 Days: A measurement option for DSM-5 Major Depressive Disorder (MDD) severity," Population Health Metrics, vol. 9, no. 1, 2011.

[16] A. BECK, "An Inventory for Measuring Depression", Archives of General Psychiatry, vol. 4, no. 6, p. 561, 1961. Available: 10.1001/archpsyc.1961.01710120031004.

[17] A. Takahashi, "Rating scale for depression", International Clinical Psychopharmacology, vol. 13, no. 1, p. 53, 1998. Available: 10.1097/00004850-199801000-00048.

[18] Clinical methods: The history, physical, and laboratory examinations. 3rd edition. LexisNexis UK, 1990.

[19] J. McCambridge, J. Witton, and D. R. Elbourne, "Systematic review of the Hawthorne Effect: New Concepts are needed to study research participation effects," Journal of Clinical Epidemiology, vol. 67, no. 3, pp. 267–277, 2014.

[20] K. Spak, "New Blood Test Checks for Depression", Newser, 2022. [Online]. Available: http://www.newser.com/story/119373/new-blood-testchecks-for-depression.html. [Accessed: 27- Sep-2022].

[21] K. Pajer et al., "Discovery of blood transcriptomic markers for depression in animal models and pilot validation in subjects with early-onset major depression", Translational Psychiatry, vol. 2, no. 4, pp. e101-e101, 2012. Available: 10.1038/tp.2012.26.

[22] P. Broca, "Broca's Speech Area", Rivista di Neuroradiologia, vol. 15, no. 3, pp. 298-299, 2002. Available: 10.1177/197140090201500321.

[23] J. Mundt, P. Snyder, M. Cannizzaro, K. Chappie and D. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology", Journal of Neurolinguistics, vol. 20, no. 1, pp. 50-64, 2007. Available: 10.1016/j.jneuroling.2006.04.001.

[24] M. Al-Mosaiwi and T. Johnstone, "In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation", Clinical Psychological Science, vol. 6, no. 4, pp. 529-542, 2018. Available: 10.1177/2167702617747074.

[25] A. Tackman et al., "Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis.", Journal of Personality and Social Psychology, vol. 116, no. 5, pp. 817-834, 2019. Available: 10.1037/pspp0000187.

[26] P. Wu, R. Wang, H. Lin, F. Zhang, J. Tu, and M. Sun, "Automatic depression recognition by Intelligent Speech Signal Processing: A Systematic Survey," CAAI Transactions on Intelligence Technology, 2022.

[27] M. Emerich and R. Lupu, "Improving speech emotion recognition using fusion and time-frequency analysis," in *2015 International Conference on Control, Decision and Information Technologies (CoDIT)*, Barcelona, 2015, pp. 563–568, doi: 10.1109/CoDIT.2015.7371891.

[28] H. Wang, Y. Liu, X. Zhen, and X. Tu, "Depression speech recognition with a three-dimensional convolutional network," Frontiers in Human Neuroscience, vol. 15, 2021.

[29] A. Biswal, "Recurrent neural network (RNN) tutorial: Types and examples [updated]: Simplilearn," Simplilearn.com, 11-Aug-2022. [Online]. Available: https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn. [Accessed: 09-Oct-2022].

[30] LSTM. .

[31] H. Cai, Z. Yuan, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, Q. Zhao, Z. Liu, Z. Yao, M. Yang, H. Peng, J. Zhu, X. Zhang, G. Gao, F. Zheng, R. Li, Z.

Guo, R. Ma, J. Yang, L. Zhang, X. Hu, Y. Li, and B. Hu, "A multi-modal open dataset for Mental-Disorder Analysis," Scientific Data, vol. 9, no. 1, 2022.

[32] E. Ciftci, H. Kaya, H. Gulec, and A. A. Salah, "The Turkish audio-visual bipolar disorder corpus," 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), 2018.

[33] L.-P. Morency, G. Stratou, D. DeVault, A. Hartholt, M. Lhommet, G. Lucas, F. Morbini, K. Georgila, S. Scherer, J. Gratch, S. Marsella, D. Traum, and A. Rizzo, "Simsensei Demonstration: A perceptive virtual human interviewer for healthcare applications," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, no. 1, 2015.

[34] J. Gratch, R. Artstein, and G. Lucas, "The Distress Analysis Interview Corpus of human and computer interviews," European Language Resources Association, vol. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014.

[35] D. France, R. Shiavi, S. Silverman, M. Silverman and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk", IEEE Transactions on Biomedical Engineering, vol. 47, no. 7, pp. 829-837, 2000. Available: 10.1109/10.846676.

[36] "," World Health Organization. [Online]. Available: https://www.who.int/zh/news-room/fact-sheets/detail/depression. [Accessed: 07-Jan-2023].

[37] F. Ringeval, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, M. Pantic, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, and S. Amiriparian, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with AI, and Cross-Cultural Affect Recognition," Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop - AVEC '19, 2019.

[38] "Data preprocessing in Data Mining," GeeksforGeeks, 29-Jun-2021. [Online]. Available: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/. [Accessed: 01-Nov-2022].

[39] Rick, "What are the five basic elements of music?," STUDIO NOTES ONLINE, 02-Oct-2022. [Online]. Available: https://studionotesonline.com/five-basic-elements-of-music/. [Accessed: 01-Nov-2022].

[40] A. Pampouchidou, O. Simantiraki, A. Fazlollahi, M. Pediaditis, D. Manousos, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias, F. Yang, and M. Tsiknakis, "Depression assessment by fusing high and low level features from audio, video, and text," Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 2016.

[41] A. Biswal, "Convolutional Neural Network tutorial [update]," Simplilearn.com, 21-Sep-2022. [Online]. Available: https://www.simplilearn.com/tutorials/deep-learning-tutorial/convolutional-neural-network [Accessed: 09-Oct-2022].

[42] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014," Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC '14, 2014.

[43] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013," Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, 2013.

[44] A. Biswal, "Convolutional Neural Network tutorial [update]," Simplilearn.com, 21-Sep-2022. [Online]. Available: https://www.simplilearn.com/tutorials/deep-learning-tutorial/convolutional-neural-network [Accessed: 09-Oct-2022].

[45] Y. Gong and C. Poellabauer, "Topic modeling based Multi-modal depression detection," Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017.

[46] D. Wu, "An audio classification approach based on machine learning," 2019 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS), 2019.

[47] N. Cummins, J. Epps, and E. Ambikairajah, "Spectro-temporal analysis of speech affected by depression and psychomotor retardation," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[48] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC '14, 2014.

[49] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC '14, 2014.

[50] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu, "Classification of depression state based on articulatory precision," Interspeech

2013, 2013.
[51] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu, "Classification of depression state based on articulatory precision," Interspeech 2013, 2013.
[52] "Partial proportional odds models and generalized ordinal logistic regression models," Applied Ordinal Logistic Regression Using Stata: From Single-Level to Multilevel Modeling, pp. 179–218, 2016.
[53] A. Jan, H. Meng, Y. F. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," IEEE Transactions on Cognitive and Developmental Systems, vol. 10, no. 3, pp. 668–680, 2018.
[54] P. Wu, R. Wang, H. Lin, F. Zhang, J. Tu, and M. Sun, "Automatic depression recognition by Intelligent Speech Signal Processing: A Systematic Survey," CAAI Transactions on Intelligence Technology, 2022.
[55] Y. Gong and C. Poellabauer, "Topic modeling based Multi-modal depression detection," Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017.
[56] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, "Detecting depression using vocal, facial and semantic communication cues," Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 2016.
[57] L. Chao, J. Tao, M. Yang, and Y. Li, "Multi Task Sequence Learning for depression scale prediction from video," 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 2015.
[58] M. Al Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," IEEE Transactions on Affective Computing, vol. 12, no. 1, pp. 262–268, 2021.
[59] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using Deep Learning Models," Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017.
[60] Haque, A., et al.: Measuring depression symptom severity from spoken language and 3D facial expressions. arXiv preprint arXiv:1811.08592 (2018)
[61] He, L., Cao, C.: Automated depression analysis using convolutional neural networks from speech. J. Biomed. Inform. 83, 103–111 (2018)
[62] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," Interspeech 2018, 2018.
[63] Z. Du, W. Li, D. Huang, and Y. Wang, "Bipolar disorder recognition via multi-scale discriminative audio temporal representation," Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, 2018.
[64] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 5688-5691, doi: 10.1109/ICASSP.2011.5947651.
[65] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using Deep Neural Network and extreme learning machine," Interspeech 2014, 2014.
[66] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.
[67] Chang-Hyun Park, Dong-Wook Lee and Kwee-Bo Sim, "Emotion recognition of speech based on RNN," Proceedings. International Conference on Machine Learning and Cybernetics, 2002, pp. 2210-2213 vol.4, doi: 10.1109/ICMLC.2002.1175432.
[68] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016," Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 2016.
[69] H. Jiang, B. Hu, Z. Liu, G. Wang, L. Zhang, X. Li, and H. Kang, "Detecting depression using an ensemble logistic regression model based on multiple speech features," Computational and Mathematical Methods in Medicine, vol. 2018, pp. 1–9, 2018.
[70] H. Jiang, B. Hu, Z. Liu, G. Wang, L. Zhang, X. Li, and H. Kang, "Detecting depression using an ensemble logistic regression model based on multiple speech features," Computational and Mathematical Methods in Medicine, vol. 2018, pp. 1–9, 2018.
[71] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017," Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017.
[72] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," Proceedings of the 25th ACM international conference on Multimedia, 2017.
[73] Y. Dong and X. Yang, "A hierarchical depression de-

tection model based on vocal and emotional cues," Neurocomputing, vol. 441, pp. 279–290, 2021.

[74] Z. Huang, J. Epps, and D. Joachim, "Exploiting vocal tract coordination using dilated CNNS for depression detection in naturalistic environments," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

[75] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017.

[76] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017.

[77] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017.

[78] H. Sun, J. Liu, S. Chai, Z. Qiu, L. Lin, X. Huang, and Y. Chen, "Multi-Modal Adaptive Fusion Transformer Network for the estimation of depression level," Sensors, vol. 21, no. 14, p. 4764, 2021.

[79] He, L., Cao, C.: Automated depression analysis using convolutional neural networks from speech. J. Biomed. Inform. 83, 103–111 (2018)

[80] "Partial proportional odds models and generalized ordinal logistic regression models," Applied Ordinal Logistic Regression Using Stata: From Single-Level to Multilevel Modeling, pp. 179–218, 2016.

[81] Z. Du, W. Li, D. Huang, and Y. Wang, "Bipolar disorder recognition via multi-scale discriminative audio temporal representation," Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, 2018.

[82] A. Jan, H. Meng, Y. F. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," IEEE Transactions on Cognitive and Developmental Systems, vol. 10, no. 3, pp. 668–680, 2018.

[83] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using Deep Learning Models," Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017.

[84] Haque, A., et al.: Measuring depression symptom severity from spoken language and 3D facial expressions. arXiv preprint arXiv:1811.08592 (2018)

[85] He, L., Cao, C.: Automated depression analysis using convolutional neural networks from speech. J. Biomed. Inform. 83, 103–111 (2018)

[86] A. Vázquez-Romero and A. Gallardo-Antolín, "Automatic detection of depression in speech using ensemble Convolutional Neural Networks," Entropy, vol. 22, no. 6, p. 688, 2020.

[87] M. Niu, B. Liu, J. Tao, and Q. Li, "A Time-frequency channel attention and vectorization network for automatic depression level prediction," Neurocomputing, vol. 450, pp. 208–218, 2021.

[88] J.-T. Chien and T.-W. Lu, "Deep recurrent regularization neural network for speech recognition," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[89] L. Chao, J. Tao, M. Yang, and Y. Li, "Multi Task Sequence Learning for depression scale prediction from video," 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 2015.

[90] M. Al Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," IEEE Transactions on Affective Computing, vol. 12, no. 1, pp. 262–268, 2021.

[91] P. Wu, R. Wang, H. Lin, F. Zhang, J. Tu, and M. Sun, "Automatic depression recognition by Intelligent Speech Signal Processing: A Systematic Survey," CAAI Transactions on Intelligence Technology, 2022.

[92] J.-T. Chien and T.-W. Lu, "Deep recurrent regularization neural network for speech recognition," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[93] "WMA - The World Medical Association-WMA Declaration of helsinki – ethical principles for medical research involving human subjects," The World Medical Association. [Online]. Available: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/. [Accessed: 05-Jan-2023].

[94] "Wecon sent - hosting premium casino games has never been so Easy," Wecon Sent - Hosting Premium Casino Games Has Never Been So Easy. [Online]. Available: http://www.weconsent.us/. [Accessed: 05-Jan-2023].

[95] M. F. Waheed and S. Bernadin, "In-Situ Analysis of Vibration and Acoustic Data in Additive Manufacturing," in SoutheastCon 2024 (SoutheastCon 2024), Atlanta, USA, pp. 5.9, Mar. 15, 2024.