International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Identification of Depression from the Speech of A Person

Rutik Deshmukh[1], Supriya Lande[2], Onkar Bhosale[3], Akshay Waghmode[4], Anupama Phakatkar[5]

[1,2,3,4]Student, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India
[5]Professor, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India

Abstract:
The world's population has experienced profound psychological effects over time due to stress, anxiety, and fast-paced modern lifestyles. Accurate mapping of the diverse forms of human biology is made possible by the global technological development in healthcare, which digitizes scopious data more than usual measuring methods. Healthcare data analysis can now be done more effectively with the use of machine learning (ML), thanks to its vast volume of data. To forecast the likelihood of mental illnesses and, consequently, carry out possible treatment outcomes, machine learning techniques are being applied to the field of mental health. This review paper outlines many machine learning techniques that are employed in the identification of depression. Classification, Deep learning, and ensemble—are used to group the machine learning-based depression detection systems. A comprehensive approach for identification depression is introduced, which includes data extraction, pre-processing, ML classifier training, detection, classification, and performance assessment. Additionally, it provides a summary that identifies the goals and constraints of various research projects that have been presented in the field of depression detection. It also covers potential directions for future research in the realm of depression diagnosis.

Index Terms - Depression Detection, Support Vector Machine(SVM), Convolutional Neural Networks (CNNs), Audio Analysis, Video Analysis, Long Short-Term Memory (LSTM), Attention Mechanisms, Ensemble Learning, Transformers.

INTRODUCTION

The contemporary lifestyle has a psychological influence on individuals, leading to emotional distress and depression. Depression is a prevalent mental condition that impacts a person's cognition and psychological growth. According to the World Health Organization (WHO), roughly 1 billion individuals are affected by mental disorders, and more than 300 million people worldwide suffer from depression. Depression often leads to suicidal ideation in individuals, with approximately 800,000 people taking their own lives each year. Therefore, addressing the mental health burden requires a comprehensive approach. Depression can have adverse effects on an individual's socioeconomic status, often causing them to withdraw from social interactions. Effective approaches to combat depression include counseling and psychological therapies. Machine learning is dedicated to creating algorithms capable of learning complex patterns independently. This ability allows them to find solutions to new problems by drawing upon previous data and solutions. The application of ML techniques in the healthcare sector has proven to be highly practical due to their capacity to process vast amounts of diverse data and provide valuable clinical insights. ML-based methods are particularly beneficial in understanding mental conditions and supporting mental health professionals in making predictive decisions. Potential to assist medical specialists and clinicians in making informed decisions regarding the most suitable diagnostic approach for individuals with depression. The paper covers the following key aspects:

1. The significance of studies that extract insights related to mental health.
2. A broad framework for diagnosing depression, encompassing databases, preprocessing, the training of ML classifiers, detection classification, and the evaluation of performance.
3. A comprehensive survey of various ML algorithms employed for diagnosing depression, categorized according to their techniques for detecting depression.
4. A discussion of the limitations encountered in the reviewed studies within the field of depression identification, aiming to enhance the understanding of ML approach selection for clinicians and healthcare professionals.
5. Exploration of potential future research avenues in the domain of depression identification.

By examining these aspects, this review paper aims to serve as a valuable resource for professionals seeking to enhance the identification of depression using ML methods.

DATABASES

Detecting deep depression necessitates ample data for training a discriminative model. Given the sensitivity surrounding depression, gathering data presents challenges. Consequently, numerous research teams have endeavored to compile their databases to investigate tools for assessing depression. In this context, we present established databases frequently utilized in reviewed studies for detecting depression. Additionally, we discuss privately released databases. Table 1 outlines the mentioned databases, encompassing subject numbers, annotation scores, availability, and further specifics.

Sr no	**Database**	**Modality**	**Subjects**	**Annotations**	**Public/Private**
1.[35]	DementiaBank(1994)	A+V+T	226	HAMD	Public
2.[36]	ORI (2009)	V	139	Manual annotation	Private
3.[37]	BlackDog (2009)	A+V	80	DSM-IV, HAMD>15	Private
4.[38]	ORYGEN (2011)	V	191	Manual Annotation	Private
5.[39]	AVEC 2013 (2013)	A+V	292	BDI-II	Public
6.[40]	AVEC 2014 (2014)	A+V	292	BDI-II	Public

Sr no	Database	Modality	Subjects	Annotations	Public/Private
7.[41]	DAIC-WOZ (2014)	A+V+T	189	PHQ-8	Public
8.[42]	CHI-MEI (2016)	V	53	DSSS, HAM-D	Private
9.[43]	Pittsburgh (2018)	A+V	57	DSM-IV, HAMD>15	Public
10.[44]	BD (2018)	A+V	46	DSM-V	Public
8.[45]	EDAIC-WOZ (2019)	A+V+T	219	PHQ-8	Public
12.[46]	MODMA (2020)	A+EEG	EEG-128(53), EEG-3(55), A(55)	PHQ-8	Public

Table1: Summary of the Audiovisual databases which have been adopted in the reviewed works for the last 20 years. Abbreviations:DPRD–Depressed,SCDL–Suicidal,NTRL–Neutral,not depressed or suicidal, M–Number of males, F–Number of Females, DSM-Diagnostic and Statistical Manual of Mental Disorders, HAMD-Hamilton Rating Scale for Depression,BDI-Beck Depression Inventory, QIDS-Quick Inventory of Depressive Symptomology, PHQ-8-Patient Health Questionnaire. Note:where DSM is present as a clinical score for all depressed patients in corpus to meet criteria for Major Depressive Disorder.

Gathering data on depression involves recruiting participants from hospitals or psychological clinics, which poses the greatest challenge in depression research. As evidenced in existing studies, depressed individuals or healthy controls are evaluated according to the DSM-IV standard and/or HAMD. The BDI is widely used for predicting the severity of depression. Alternatively, various criteria such as PHQ-8 and BDI-II are employed to assess depression-related symptoms. Additionally, several studies have employed alternative recruitment methods, including flyers, posters, social networks, personal contacts, and mailing lists.

RELATED WORK

Since the 1980s, the field of speech emotion recognition has witnessed significant developments. Early research by Bezooijen [1] and Tolkmitt [2] in the 1980s marked the beginning of using acoustic statistical features to identify emotions. In 1999, Moriyama [3] introduced the concept of associating speech with emotions and applied it to an e-commerce system, enabling the collection of user speech and emotion recognition. Early studies on speech emotion and depression detection primarily involved various machine learning techniques such as Gaussian Mixtures Models (GMMs), Hidden Markov Models (HMMs), Support Vector Machine (SVM), and Random Forest (RF).

The general approach in these traditional machine learning methods was to extract relevant features from speech data and then utilize machine learning algorithms to study the relationship between these features and the degree of depression. However, the crucial aspect in traditional machine learning approaches was the feature selection process, which directly impacted the accuracy of depression recognition. While these methods had the advantage of not requiring extensive amounts of data for training, they suffered from the challenge of determining the quality of features, potentially leading to the omission of critical features and subsequently reducing the accuracy of depression identification.

Recent advancements in the field of speech emotion recognition have incorporated deep learning techniques such as deep neural networks (DNNs) and convolutional neural networks (CNNs), which have shown promising results in improving accuracy. Deep learning technology offers the advantage of extracting high-level semantic features. Meyer and his colleagues[4] introduced Deep Neural Networks (DNNs) for the purpose of speech emotion recognition. Another study by Han et al. [5] presented a speech emotion classification system utilizing a combination of DNN and Extreme Learning Machine (ELM). In this approach, traditional acoustic speech features were inputted into the DNN, leading to the generation of a probability distribution of segmental emotional states. Subsequently, utterance-level features were constructed, and ELM was employed for classification. Le Yang [6] took inputs as the traditional hand-crafted speech feature set as preliminary features into a deep convolutional network to learn and predict the depression level of depressed patients.

However, a limitation of DNNs is their reliance on personalized features as input which can be influenced by various factors such as speech styles, speech content, and context. This limitation hinders their application in real-world settings that involve different speakers. To address this issue, Bertero and Fung [7] applied Convolutional Neural Networks (CNN), a technology predominantly used in image processing, to speech emotion recognition, yielding positive results. Nonetheless, the CNN used was a relatively simple shallow model and failed to fully leverage the temporal correlation characteristics of speech. S. Dhams' [8]study explores automated depression detection using a multimodal approach, analyzing audio, video data from the DAIC-WOZ dataset. Machine learning techniques, including SVM, GMM, CNN and Decision level fusion, are employed for classification. The results obtained were able to cross the provided baseline on validation data set by 17% on audio features and 24.5% on video features.

Recognizing the strength of Recurrent Neural Networks (RNNs) in handling timing-related problems, Park et al. [9] incorporated RNNs into speech emotion recognition. Subsequently, researchers improved RNNs and introduced models like LSTM (Long Short-Term Memory), GRU, and QRNN [10]. K. Cho enhances their capabilities. However, one of the major drawbacks of RNNs is their difficulty in training and their susceptibility to overfitting issues, especially when dealing with small-scale emotional datasets. To address these challenges, some variants aimed to combine the advantages of CNNs and RNNs in a model known as Convolutional Recurrent Neural Network (CRNN) for speech emotion recognition. In this approach,[11] low dimensional features of speech served as the foundational features for speech emotion feature extraction. The CNN was employed to map these features and the LSTM was used to capture sentence level timing information.. Chao et al.[12] introduce a multimodal depression prediction model using audiovisual input, employing LSTM-RNN to encode temporal dynamics in abnormal audio-visual behavior and multi-task learning to incorporate emotion information, facilitating concurrent learning of abnormal behavior detection and emotion recognition. Ma & Yang [13] address data representation and sample imbalance issues through DepAudioNet, combining CNN and LSTM for comprehensive audio representation, and introducing a random sampling strategy during training to balance positive and negative samples, demonstrating effectiveness on the DAIC-WOZ dataset. Mohamad J, and Guodong G. [14] focus on predicting depression levels using visual data, employing Convolutional

3D (C3D) models and RNNs to capture spatiotemporal features from videos, Pre-trained C3D models were fine-tuned, and RNNs were used to enhance results by learning from temporal feature sequences. showcasing the potential of deep learning models for accurate depression level prediction through video analysis. The research was focused on AVEC 2013 and AVEC 2014 datasets. The research demonstrated a significant accuracy of approximately 85.7% using visual data. Lin & Chen [22] was focused on automated depression detection using speech signals and linguistic content. It employed deep learning techniques, including Bidirectional Long Short-Term Memory (BiLSTM) and One-Dimensional Convolutional Neural Network (1D CNN), to achieve state-of-the-art results. These studies collectively highlight the potential of these methodologies and algorithms and multimodal approaches in addressing challenges associated with depression prediction and emotion recognition.

Haque et al. [15] employed a multimodal methodology to assess depression symptom severity by analyzing both 3D facial expressions and spoken language. Their approach integrated sentence-level embeddings and causal convolutional networks (C-CNNs)[34]. Leveraging the DAIC-WOZ dataset, their model achieved a notable accuracy of 76.9% with an average error rate of 3.67 points on the PHQ scale for predicting depression symptom severity.

Adrian R.[16] proposes a method in the Depression Classification Sub-Challenge (DCC) at AVEC-2016. The approach involves preprocessing speech files into log spectrogram sequences and ensuring a balance between positive and negative samples. For classification, One-Dimensional CNNs are employed, with multiple models trained using distinct initialization. The Ensemble Averaging technique is used to combine individual predictions for each speaker, resulting in superior performance compared to SVM-based systems, CNN+LSTM systems (like DepAudionet), and single CNN based classifiers in depression detection.

Lam et al. [17] introduced an innovative methodology including 1D convolutional neural networks (CNNs) and transformers, along with data augmentation via topic modeling. Their study, using the DAIC-WoZ dataset, achieved significant advancements in depression detection accuracy, surpassing previous state-of-the-art methods with higher F1 scores. The audio-only model achieved 67% accuracy, while the text-only model achieved 78%. Notably, the multi-modal fusion model achieved an impressive accuracy of 87%. The study underscores the effectiveness of integrating context-aware and data-driven methodologies and highlights the potential impact on advancing AI-based tools for mental health disorders. Future research directions include enhancing multi-modal fusion techniques to further improve model accuracy. [18] and [27] incorporated acoustic words (AW) and landmark words (LW) in speech. By leveraging token-based features and embedding techniques such as LDA, Word2Vec, and GloVe, the study evaluated datasets including DAIC-WOZ and SH2-FS. It primarily employed Linear Support Vector Machines (SVM) as the machine learning algorithm Results showcased the superiority of token-based methods over conventional frame-based features, with the hybrid approaches notably enhancing detection accuracy. Particularly, the fusion of AW and LW using hybrid methods achieved an F1 score of up to 0.667 for depression detection.

Yang & Le [20] introduces a pioneering approach leveraging Deep Convolutional Generative Adversarial Networks (DCGAN) to enhance depression severity estimation via speech analysis. Speech acoustic features are translated into 2D image-like representations, and DCGAN models are employed for feature generation, guided by depression intensity constraints. Utilizing the AVEC2016 depression dataset, the experiments showcase exceptional accuracy, achieving a significant reduction in root mean square error (RMSE) to 5.520 and mean absolute error (MAE) to 4.634.

Madhavi and her team [21] delve into detection of work-related stress from audio streams in cyber-physical environments, addressing their unique challenges. It involves preprocessing audio streams to extract low-level features using Short-term Fourier Transform, followed by high-level feature extraction through a Convolutional Neural Network (CNN) and Growing Self-Organizing Map (GSOM) algorithm. The study utilizes DAIC-WOZ dataset. Techniques like noise removal, pitch, and speed augmentation are employed to enhance data quality and address imbalances. The CNN-GSOM model demonstrates promising performance in distinguishing distressed and non-distressed individuals, performing existing methods with F1 scores of 82% and 64% for normal and distressed classes, respectively. Insights from the paper suggest future research should explore further refinements of the proposed approach, scalability to larger datasets, and integration with real-time monitoring systems for proactive stress management in cyber-physical environments. Additionally, investigating the generalization of the model across diverse industrial contexts and exploring multimodal approaches for comprehensive stress assessment could be promising avenues for future exploration.

Wang [23] evaluated depression detection algorithms using speech data on DAIC-ori and DAIC-mute-removed datasets. The 3D-CBHGA model consistently outperformed SVM, RF, and 1D-CBBG, demonstrating its effectiveness. The research highlighted the impact of silence in speech on depression recognition. The key technologies and algorithms used included 3D-CBGHA, SVM (Support Vector Machines), RF (Random Forest), and 1D-CBBG (not explicitly defined but used for comparison). These methods were tested to assess the performance in detecting depression through speech analysis.

S. H. Dumpala & team [24] explored depression severity estimation from speech recordings by integrating acoustic features and sentiment/emotion embeddings. They utilized FORBOW dataset comprising recordings from 526 subjects, the study analyzes speech segments annotated by clinical experts for sentiment and emotion cues. Methodologically, it extracts spectral and excitation source-based features and employs Multi-task-CNNs (MT-CNNs) and Sentiment-Emotion Embeddings (MT-CNN-SE) for regression and classification tasks, optimizing with dropout rates and the Adam optimizer. The models exhibit promising accuracy in depression severity estimation, particularly with the integration of combined spectral and prosodic features and sentiment/emotion embeddings. The study concludes by highlighting clinical implications and suggesting future research avenues to further refine accuracy and applicability in clinical settings.

B. Maji & M. [25] presented the study presenting a Speech Emotion Recognition (SER) model using Convolutional Capsule (Conv-Cap) and Bi-directional Gated Recurrent Unit (Bi-GRU) networks, handling varying speech lengths. It leveraged a dual-channel self-attention mechanism and extracted six spectral features, optimizing selection through self-attention layers. A confidence-based fusion method enhanced classification, resulting in high recognition rates. The model excelled in SER performance, outperforming traditional machine learning algorithms, with weighted and unweighted accuracies reaching 90.31% and 87.61%, respectively. Datasets IEMOCAP, EMO-DB, and Odia were used in comprehensive comparisons, demonstrating the model’s effectiveness. Technologies applied include ConvCap networks, Bi-GRU networks, and self-attention layers.

Yongming Huang [26] was focused on diagnosing depression through speech signals in a non-invasive and cost-effective manner. It introduced a hierarchical attention temporal convolutional network (HATCN) model, incorporating attention mechanisms for feature extraction. This study addressed sample imbalance using a periodic focal loss function. It utilized technologies such as acoustic signal processing, speech processing, and voice communication analysis to achieve improved depression recognition.

In [27] Marouane Birjali explored machine learning and semantic sentiment analysis techniques for predicting suicide sentiments in social networks. Various models, including support vector machines (SVM), random forest, and deep learning neural networks, were utilized, with SVM achieving the highest accuracy of 87%, followed by random forest with 82%, and the neural network with 78%. The experiments were conducted on a large-scale dataset comprising social media posts containing sentiments related to suicide, showing promising outcomes for accurately predicting suicidal sentiments and enabling early detection and intervention. Similarly in [28] Mandar & V. focused on testing a classifier for depression detection using Twitter data. Multinomial Naive Bayes achieved a higher F1 score (83.29) compared to SVM (79.73), with corresponding accuracies of 83% and 79%, respectively. Despite the effectiveness of text-based emotion AI in detecting depression, challenges arose from non-standard text in tweets, impacting classifier accuracy. The paper recommends future research to address these challenges and proposes adding an expert-based layer to reduce false positives.

Daun Shin et al. [29] employed a rigorous methodology combining structured psychiatric interviews, voice recordings, feature extraction, and machine learning algorithms. 93 subjects are recruited and categorized into not depressed, minor depressive episode, and major depressive episode groups based on psychiatric evaluations. Voice recordings are obtained during Mini-International Neuropsychiatric Interview (MINI) sessions, with voice features extracted from the recordings. Various machine learning algorithms, including logistic regression, Gaussian Naive Bayes, SVM and multilayer perceptron, are employed to analyze the data. The multilayer perceptron emerges as the most accurate predictor, achieving an area under the curve (AUC) of 0.79 and 0.58 for minor and major depressive episodes, respectively, with precision averaging 65.6% and recall averaging 66.2%. These findings suggest promising avenues for utilizing voice analysis as a diagnostic tool for mental health assessment. However, further research with larger sample sizes and longitudinal studies is essential to validate these results and explore integration with other modalities.

Flaming Yin et al. [30] introduced a transformer-CNN-CNN (TCC) approach for depression detection in speech, utilizing the DAIC-WOZ and MODMA datasets. The TCC model, employing a parallel CNN and a transformer in a three-stream parallel structure, demonstrated superior performance over single CNN, parallel CNN, and transformer models. Experimental results revealed that the TCC model, particularly with the TCC-kernel setup, achieved a remarkable F1-score of 96.7% on the MODMA dataset, surpassing existing models like 2D-CNN-LSTM and CNN-BLSTM with self-attention. The TCC approach excelled in precision, recall, and F1-score, establishing its effectiveness in depression detection. Notably, the TCC-softmax setup exhibited significant performance improvements, highlighting the proposed model's superiority.

Feifan W. and X. Shen [31] proposed a CNN–LSTM neural network for speech emotion recognition using the RAVDESS database. Features extracted include Mel-frequency cepstral coefficients (MFCC), inverted MFCC (IMFCC), and Teager energy operator coefficients (TEOC). Through feature ablation experiments, IMFCC emerged as more effective in capturing emotional features than MFCC alone. Fusion of TEOC and IMFCC further enhanced recognition accuracy. Comparative analysis with models like Self-Supervised Learning and Transformer indicated superior performance in terms of network structure, parameter transfer, and training duration. The method excelled in recognizing emotions such as surprise, neutral, anger, and calm, presenting practical applications in domains like intelligent assisted driving.

Sri Harsha Dumpala [32] and colleagues explored the relationship between depression manifestations in speech and features commonly utilized for speaker recognition. Employing a blend of acoustic and linguistic features extracted from speech data, the research aimed to identify signs of depression. Machine learning models such as SVM and DNN were used to carry out the research. The experiments demonstrated that SVM achieved an accuracy of 78%, while DNN achieved 82% accuracy in distinguishing between depressed and non-depressed individuals. DAIC-WoZ and Vocal Mind Datasets were utilized and the results revealed substantial overlaps between features used for depression detection and speaker recognition, underscoring the potential of speech-based biomarkers in identifying individuals at risk of depression.

A. Y. Kim et al [33] explored automatic depression detection through speech analysis, particularly focusing on a large-scale assessment using acoustic characteristics in the Korean language which was conducted on 153 patients with major depressive disorder and 165 healthy controls, the research evaluated three approaches: conventional machine learning models based on acoustic features, a proposed deep convolutional neural network (CNN) model, and models using pretrained networks. The results indicated that the CNN model achieved an accuracy of 78.14%, outperforming conventional and pretrained models. The study highlighted the potential of smartphone-based, easily accessible methods for automatic depression identification, emphasizing the significance of speech data analysis in predicting depressive states. The study investigated automatic depression detection through speech analysis, employing text-dependent read speech tasks in Korean. Emphasizing the benefits of controlled data acquisition, the findings underscored the potential of this method for private and accurate analysis of acoustic patterns linked to depression. While acknowledging study limitations, such as small sample sizes, the research opened promising avenues for widespread depression screening in everyday life.

These diverse approaches reflect the evolving landscape of speech emotion recognition, highlighting the importance of addressing personalized features, temporal characteristics, and multi modal data integration for improved accuracy and practical applications in various contexts.

CONCLUSION

The rapid evolution of machine learning (ML) techniques has paved the way for significant advancements in the identification and diagnosis of depression. The reviewed literature underscores the transformative impact of these technologies on mental health care, offering innovative solutions to address the complex challenges posed by stress, anxiety, and fast-paced modern lifestyles.

The diverse range of ML algorithms discussed, from traditional methods to sophisticated deep learning models, reflects the dynamic landscape of depression detection. The integration of multimodal data sources, including audio and video analysis, demonstrates the commitment to holistic approaches that capture the nuanced manifestations of mental health conditions. By emphasizing the significance of comprehensive approaches, the paper aligns with the growing recognition that mental health is a multifaceted domain requiring nuanced strategies for accurate assessment and treatment.

The survey not only highlights the successes and promising results achieved by various ML methodologies but also candidly addresses the limitations and challenges encountered in existing studies. The paper's exploration of deep learning models, including CNNs, LSTM, CRNN, TCC, HATCN, MT-CNNs,1D-CBBG, 3D-CBGHA, Transformers etc showcases the ongoing pursuit of improved accuracy and robustness in the face of personalized features, temporal characteristics, and diverse data modalities.

As technology continues to intersect with mental health care, the discussed studies point towards a future where innovations such as speech recognition, computer vision, and natural language processing converge to offer low-cost, universally accessible mental health support. The potential deployment of such technologies on widely available devices like smartphones underscores the potential for transformative impacts on global mental health outcomes.

In summary, this survey paper serves as a comprehensive resource for healthcare professionals, researchers, and clinicians seeking to leverage ML techniques for depression identification. The integration of insights from various studies positions this review as a guiding document for the ongoing quest to refine and enhance depression detection methodologies, ultimately contributing to improved mental health care on a global scale.

REFERENCES

1. R. van Bezooijen, S. A. Otto, and T. A. Heenan, "Recognition of vocal expressions of emotion,"Journal of Cross-Cultural Psychology, vol. 14, pp. 387 – 406, 1983.
2. F. J. Tolkmitt and K. R. Scherer, "Effect of experimentally induced stress on vocal parameters.," Journal of Experimental Psychology: Human Perception and Performance, vol. 12, no. 3, p. 302, 1986.
3. T. Moriyama and S. Ozawa, "Emotion recognition and synthesis system on speech," in Proceedings IEEE International Conference on Multimedia Computing and Systems, vol. 1, pp. 840–844, IEEE, 1999.
4. A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5688–5691, IEEE, 2011.
5. K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in Interspeech 2014, 2014.
6. L. Yang, D. Jiang, W. Han, and H. Sahli, "Dcnn and dnn based multimodal depression recognition," in 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII),pp. 484–489, IEEE, 2017.
7. D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5115–5119, IEEE, 2017.
8. S. Dham, A. Sharma, and A. Dhall, "Depression scale recognition from audio, visual and text analysis," arXiv preprint arXiv:1709.05865, 2017.
9. C.-H. Park, D.-W. Lee, and K.-B. Sim, "Emotion recognition of speech based on rnn," in Proceedings. International Conference on Machine Learning and Cybernetics, vol. 4, pp. 2210–2213, IEEE, 2002.
10. K. Cho, B. Van Merrienboer,¨ D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," arXiv preprint arXiv:1409.1259, 2014.
11. S. Basu, J. Chakraborty, and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," in 2017 2nd International Conference on Communication and Electronics Systems (ICCES), pp. 333–336, IEEE, 2017.
12. L. Chao, J. Tao, M. Yang, and Y. Li, "Multi task sequence learning for depression scale prediction from video," in 2015 International conference on affective computing and intelligent interaction (ACII),pp. 526–531, IEEE, 2015.
13. X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in Proceedings of the 6th international workshop on audio/visual emotion challenge, pp. 35–42, 2016.
14. M. AlJazaery 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, 2018.

15. A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei, "Measuring depression symptom severity from spoken language and 3d facial expressions," arXiv preprint arXiv:1811.08592, 2018.

16. A. Vazquez´-Romero and A. Gallardo-Antol´ın, "Automatic detection of depression in speech using ensemble convolutional neural networks," Entropy, vol. 22, no. 6, p. 688, 2020.

17. G. Lam, H. Dongyan, and W. Lin, "Context-aware deep learning for multi-modal depression detection," in ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 3946–3950, IEEE, 2019.

18. Z. Huang, J. Epps, D. Joachim, and V. Sethu, "Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2, pp. 435–448, 2019.

19. L. Yang, D. Jiang, and H. Sahli, "Feature augmenting networks for improving depression severity estimation from speech signals," IEEE Access, vol. 8, pp. 24033–24045, 2020.

20. I. Madhavi, S. Chamishka, R. Nawaratne, V. Nanayakkara, D. Ala-hakoon, and D. De Silva, "A deep learning approach for work related stress detection from audio streams in cyber physical environments," in 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), vol. 1, pp. 929–936, IEEE, 2020.

21. L. Lin, X. Chen, Y. Shen, and L. Zhang, "Towards automatic depression detection: A bilstm/1d cnn-based model," Applied Sciences, vol. 10, no. 23, p. 8701, 2020.

22. H. Wang, Y. Liu, X. Zhen, and X. Tu, "Depression speech recognition with a three-dimensional convolutional network," Frontiers in human neuroscience, vol. 15, p. 713823, 2021.

23. S. H. Dumpala, S. Rempel, K. Dikaios, M. Sajjadian, R. Uher, and S. Oore, "Estimating severity of depression from acoustic features and embeddings of natural speech," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7278–7282, IEEE, 2021.

24. B. Maji and M. Swain, "Advanced fusion-based speech emotion recogni-tion system using a dual-attention mechanism with conv-caps and bi-gru features," Electronics, vol. 11, no. 9, p. 1328, 2022.

25. Y. Huang, Y. Ma, J. Xiao, W. Liu, and G. Zhang, "Identification of depression state based on multi-scale acoustic features in interrogation environment," IET Signal Processing, vol. 17, no. 4, p. e12207, 2023.

26. Marouane Birjali, M Erritali, A.B hassane "Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks"65–721877-0509,IEEE,2017.

27. Mandar Deshpande, Vignesh Rao "Depression Detection using Emotion Artificial Intelligence" 978-1-5386-1959-9/17/$31.00, MDPI,2023.

28. Daun Shin,Won Ik Cho,C. Hyung Keun Park,Sang Jin Rhee,Min Ji Kim,Hyunju Lee, Nam Soo Kim, and Yong Min Ahn "Detection of Minor and Major Depression through Voice as a Biomarker Using Machine Learning" 10.1007/s10068-020-00809-4,IEEE,2021.

29. Flaming Yin,Jing Du ,Xinzhou Xu ,and Li Zhao "Depression Detection in Speech Using Transformer and Parallel Convolutional Neural Networks",2079-9292/12/2/328,MDPI,2023

30. Feifan WangORCID and Xizhong Shen"Research on Speech Emotion Recognition Based on Teager Energy Operator Coefficients and Inverted MFCC Feature Fusion"12(17),3599,MDPI,2023.

31. Sri Harsha Dumpala, Katerina Dikaios, Sebastian Rodriguez, Ross Langley, Sheri Rempel, Rudolf Uher & Sageev Oore "Manifestation of depression in speech overlaps with characteristics used to

represent and recognize speaker identity" 13:11155,IEEE,2023.

32. Ah Young Kim,Eun Hye Jang,Seung-Hwan Lee,Kwang-Yeon Choi,Jeon Gue Park,and Hyun-Chul Shin "Automatic Depression Detection Using Smartphone-Based Text-Dependent Speech Signals: Deep Convolutional Neural Network Approach"10.2196/34474,IEEE,2023.
33. S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv, 2018.
34. J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, K. L. McGonigle, The natural history of cognitive decline in Alzheimer's disease, Archives of Neurology 51 (6) (1994) 585–594.
35. N. C. Maddage, R. Senaratne, L.-S. A. Low, M. Lech, N. Allen, Video-based detection of the clinical depression in adolescents, in: En gineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, IEEE, IEEE, Minneapolis, MN, USA, 2009, pp. 3723–3726.
36. S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, G. Parker, From joyous to clinically depressed: Mood detection using spontaneous speech, in: Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, AAAI, Marco Island, Florida, 2012, pp. 141–146.
37. K. Ooi, L. Low, M. Lech, N. Allen, Prediction of clinical depression in adolescents using facial image analysis, in: WIAMIS 2011: 12th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS, WIAMIS, Delft, The Netherlands, 2011, pp. 1–4.
38. M.Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, M. Pantic, Avec2013: the continuous audio/visual emotion and depression recognition challenge, in: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, 2013, pp. 3–10.
39. M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, M. Pantic, AVEC 2014: 3D dimensional affect and depression recognition challenge, in: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, ACM, Orlando, Florida, USA, 2014, pp. 3–10.
40. J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al., The distress analysis interview corpus of human and computer interviews., in: LREC, 2014, pp. 3123–3128
41. K.-Y. Huang, C.-H. Wu, Y.-T. Kuo, F.-L. Jang, Unipolar depression vs. bipolar disorder: An elicitation-based approach to short-term detection of mood disorder., in: INTERSPEECH, 2016, pp. 1452–1456.
42. H. Dibeklio˘ glu, Z. Hammal, J. F. Cohn, Dynamic multimodal measurement of depression severity using deep autoencoding, IEEE Journal of Biomedical and Health Informatics 22 (2) (2018) 525–536.
43. E. Çiftçi, H. Kaya, H. Güleç, A. A. Salah, The turkish audio-visual bipolar disorder corpus, in: 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), IEEE, 2018, pp. 1–6.
44. Ringeval, Fabien, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt et al. "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition," in Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop, pp. 3-12. ACM, 2019.
45. H. Cai, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, Q. Zhao, et al., Modma dataset: a multi-model open dataset for mental-disorder analysis, arXiv preprint arXiv:2002.09283 (2020).