Enhanced Generalizability for Audio-Based Mental Health Detection Across Diverse Populations

Pavan Kumar R – 1GA22AI032 **Drushya K B – 1GA22AI014**

Sl. No.	Paper Title	Publication	Year	Objective	Dataset	Methodology	Result Discussion	Review	Bibliography
1	Multimodal Depression Detection: A Survey	IIT Bombay	N/A	To analyze audio-based features, emphasizing speech patterns and acoustic markers for depression detection	DAIC-WOZ, EATC, AVEC	Prosodic and spectral feature analysis, deep CNNs, Wav2vec 2.0 for feature representation	Higher accuracy with models like CNNs and Wav2vec 2.0. Challenges include noise and dataset quality	Strength: Accurate for subtle emotional differences. Weakness: Noisy interference and need for high-quality datasets	Moon, Palash, and Pushpak Bhattacharyya. Multimodal Depression Detection: A Survey. IIT Bombay
2	A Comprehensive Review of Predictive Analytics Models for Mental Illness Using ML Algorithms	Healthcare Analytics (Elsevier)	2024	Review machine learning approaches for early mental illness detection and propose combining modalities like social media, wearable sensors, and audio cues	Depression Audio Dataset, PAIR Dataset	Prosodic and spectral feature extraction with deep learning models (RNNs, BGRUs) and preprocessing of audio data	Deep learning models achieved up to 96.51% accuracy. Limitations include dataset dependency and audio noise	Strength: Speech intonation analysis strong. Weakness: Reliance on small datasets and quiet environments	Islam, Md. Monirul, et al. "A Comprehensive Review of Predictive Analytics Models for Mental Illness Using ML Algorithms." Healthcare Analytics, vol. 6, 2024, p. 100350
3	Investigating Generalizability of Speech-based Suicidal Ideation Detection Using Mobile Phones	ACM Interactive, Mobile, Wearable, and Ubiquitous Tech	2023	Investigate generalizability across datasets and improve model performance through unsupervised and semi-supervised domain adaptation	MDD, AVH, PT, StudentSADD datasets	Feature extraction (MFCCs), deep learning models (VGGish + LSTMs), Semi-supervised domain adaptation (SSDA), and S3 for data sub-selection	Poor cross-dataset generalization. SSDA + S3 enhanced performance by selecting optimal data subsets without labels	Strength: Rigorous cross-dataset validation. Weakness: Dataset variability impacts generalization	Pillai, Arvind, et al. "Investigating Generalizability of Speech-based Suicidal Ideation Detection Using Mobile Phones." ACM Interactive, Mobile, Wearable, and Ubiquitous Tech, vol. 7, no. 4, 2023
4	Use of ML Algorithms Based on Text, Audio, and Video Data in Predicting Anxiety and PTSD	Biological Psychiatry (Elsevier)	2024	Systematic review of ML algorithms for predicting anxiety and PTSD using behavioral data, including audio, text, and video	Emo-DB and curated speech datasets	Multimodal methods (audio + video), prosodic and spectral feature extraction, and neural network training for prediction	Audio-based studies showed high predictive power (89%). Limited representation of audio in studies	Strength: Promising results for PTSD detection. Weakness: Insufficient focus on audio	Ciharova, Marketa, et al. "Application of ML Algorithms Using Text, Audio, and Video Data for Anxiety and PTSD Predictions." Biological Psychiatry, vol. 96, no. 5, 2024. doi:10.1016/j.biopsych.2024.06.002
5	Enhancing Accuracy and Privacy in Speech-Based Depression Detection	Computer Speech and Language (Elsevier)	2024	Improve accuracy and privacy in depression detection by disentangling speaker identity from depression features	DAIC-WOZ, EATD Corpus	Speaker disentanglement using adversarial training, loss equalization, CNN-LSTM, and ECAPA-TDNN models with spectral and prosodic features	F1 score improved to 80% with LECE + CNN-LSTM. High privacy with speaker anonymization (DeID: 85%)	Strength: Innovative privacy solutions. Weakness: Limited domain-specific optimization for prosodic features	Ravi, Vijay, et al. "Enhancing Accuracy and Privacy in Speech-Based Depression Detection." Computer Speech and Language, vol. 86, 2024. doi:10.1016/j.csl.2023.101605
6	Multitask Representation Learning for Multimodal Estimation of Depression Level	IEEE Intelligent Systems	2019	Propose a multitask learning model combining acoustic, textual, and visual data for depression level regression and classification	DAIC-WOZ	Multimodal fusion using LSTM networks and deep neural networks (DNNs), attention mechanisms for relevant modality fusion	Multimodal fusion improved regression accuracy (4.93% RMSE). Audio-focused models expected to improve precision	Strength: Effective multimodal fusion. Weakness: Less emphasis on individual modalities	Qureshi, Syed Arbaaz, et al. "Multitask Representation Learning for Multimodal Estimation of Depression Level." IEEE Intelligent Systems, vol. 34, no. 5, 2019. doi:10.1109/MIS.2019.2925204
7	Multimodal Sensing for Depression Risk Detection	Sensors (MDPI)	2024	Framework integrating audio, video, and text data (AVTF-TBN) to overcome inefficiencies in traditional depression detection systems	Dataset of 1911 subjects with PHQ-9 labels	AVTF-TBN: Audio (MFCC + GRU), Video (ResNet34 + Swin Transformer), Text (BERT), and multimodal fusion with attention and residual connections	F1 score of 0.78. Multimodal fusion outperformed unimodal methods. Video data consistently showed higher performance	Strength: Well-designed multimodal fusion. Weakness: Small datasets for extreme cases	Zhang, Z., et al. "Multimodal Sensing for Depression Risk Detection: Integrating Audio, Video, and Text Data." Sensors, vol. 24, 2024. doi:10.3390/s24123714
8	TAMFN: Time-Aware Attention Multimodal Fusion Network	IEEE Transactions on Neural Systems and Rehabilitation Engineering	2023	Propose TAMFN model for efficient depression detection using non-verbal cues from audio and visual data in vlogs	D-Vlog dataset of 961 vlog videos labeled with PHQ-9 scores	Global Temporal Convolutional Network (GTCN) for dependencies, Intermodal Feature Extraction (IFE), TAMF for temporal weights, and OpenSmile for acoustic features	Weighted F1 score: 0.6582. Audio features showed higher quality, with multimodal fusion significantly improving performance	Strength: Strong temporal modeling. Weakness: Challenges with noisy and varied vlog data	Zhou, Li, et al. "TAMFN: Time-Aware Attention Multimodal Fusion Network for Depression Detection." IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, 2023. doi:10.1109/TNSRE.2022.3224135
9	MHA: A Multimodal Hierarchical Attention Model for Depression Detection in Social Media	Health Information Science and Systems	2023	Propose a multimodal hierarchical attention model using text, image, and auxiliary data for detecting depressive tendencies in social media users	Custom Sina Weibo dataset	Feature extraction with OpenSmile, within-modal attention for audio, cross-modal attention for fusion with text and images, normalization to address biases	Audio data contributed significantly to emotion detection, particularly when fused with textual data. Attention mechanisms enhanced subtle feature integration	Strength: Robust multimodal feature integration. Weakness: Dataset lacks audio, though methods are relevant for audio-based approaches	Li, Zepeng, et al. "MHA: A Multimodal Hierarchical Attention Model for Depression Detection in Social Media." Health Information Science and Systems, vol. 11, no. 6, 2023. doi:10.1007/s13755-022-00197-5
10	Multi-Explainable TemporalNet	CVPR Workshops (IEEE)	2024	Develop interpretable multimodal model for depression detection using temporal information alongside text and image features	Twitter and Reddit datasets	Multi-Explainable TemporalNet (METN): Temporal Convolutional Network with EmoBERTa (text), CLIP + DINO (images), and attention maps for interpretability	F1 scores: Twitter (0.945), Reddit (0.913). Superior temporal dependency handling and interpretability through attention maps	Strength: Effective integration of multimodal and temporal features. Weakness: Dataset demographic bias and dependence on quality data	Zafar, Anas, et al. "Multi-Explainable TemporalNet: An Interpretable Multimodal Approach Using Temporal Convolutional Network for User-Level Depression Detection." CVPRW, 2024. doi:10.1109/CVPRW63382.2024.00231