Detecting Stress Anxiety and Depression from voice tone and text responses

"Author, "Monika K , Palak Dattatraya Kota, Prathyusha N B, Purabh Pradee Singh

Department of Information Science and Engineering , NMIT

Abstract*: The global burden of mental health disorders—particularly stress, anxiety, and depression—has surged in recent years, further exacerbated by the isolation, uncertainty, and lifestyle changes imposed by the COVID-19 pandemic. Despite the growing need for psychological support, clinical diagnostics remain inaccessible to a significant portion of the population due to stigma, cost, and lack of resources. As such, scalable and objective screening mechanisms are imperative.*

This study presents a machine learning-based approach to detect psychological symptoms using voice tone and text response data. The methodology includes the alignment of audio and transcript data, extraction of statistical features from speech (using FORMANT and COVAREP toolkits), and the conversion of text into semantic vectors using a transformer-based sentence encoder (all-MiniLM-L6-v2). For labeling, a zero-shot classification model (facebook/bart-large-mnli) is employed to assign utterances into symptom categories, which are subsequently binarized. The final dataset, consisting of approximately 4,700 utterances with over 860 features, is fed into an XGBoost classifier.

Evaluation using classification metrics demonstrates robust model performance with high accuracy and interpretability. The results underline the potential of integrating acoustic and linguistic cues into an effective early warning system for mental health assessment. The system's design allows it to be scalable and deployable in real-world digital health platforms.

Keywords*: Mental Health Detection, Stress, Anxiety, Depression, Speech Analysis, NLP, XGBoost, Zero-shot Learning*

1. Introduction :

Mental health conditions represent one of the most pressing challenges in modern society, with their impact affecting both individual well-being and economic productivity. According to the World Health Organization (WHO), depression alone affects more than 264 million people globally. The social stigma attached to seeking psychiatric care, combined with a shortage of clinical professionals in many regions, further exacerbates the crisis.

In addition to social and infrastructural barriers, existing mental health diagnostic processes are often subjective, time-consuming, and require extensive professional involvement. These limitations highlight the necessity for digital solutions that are automated, efficient, and accessible. With the widespread adoption of smartphones and voice-enabled devices, there is an unprecedented opportunity to incorporate mental health assessments into everyday interactions. As individuals naturally communicate through speech and writing, these mediums offer valuable signals that can be harnessed for psychological screening.

Voice and text, as natural modes of communication, contain rich information that can be indicative of an individual's mental health status. Reduced speech energy, monotonous intonation, and lexical markers of hopelessness are well-documented correlates of psychological distress. Through modern machine learning, these subtle signals can be quantified and analyzed systematically to infer mental states in real time.

This research proposes a machine learning-based framework for detecting stress, anxiety, and depression by integrating multimodal inputs—specifically vocal tone and textual responses. Unlike many black-box deep learning approaches, our method uses XGBoost—a transparent and interpretable gradient-boosting algorithm that not only performs competitively but also provides explainability through feature importance visualization. The objective of this study is twofold: to develop a model that achieves high predictive accuracy and to ensure that the model's outputs can be interpreted and trusted in clinical or remote health settings. By fusing audio and text modalities and applying zero-shot labeling with pretrained language models, we provide a scalable and explainable approach to automated mental health assessment.

2. Literature review:

The intersection of artificial intelligence and mental health has garnered considerable attention in recent years. Researchers have explored various modalities such as text, audio, video, and physiological signals to detect psychological disorders. Among them, speech and text stand out for their non-invasiveness and ease of collection. Following are some among the several studies have been conducted in this realm:

Prediction of Depression, Anxiety and Stress Levels Using DASS-42 [1] : This study applies machine learning models, specifically SVM and Logistic Regression, to predict depression, anxiety, and stress levels from DASS-42 questionnaire data. Logistic Regression achieved the highest accuracy, highlighting the potential of ML for early mental health diagnosis.

Exploring the Effectiveness of Advanced Machine Learning Models in Speech Emotion Recognition [2] : This comparative study evaluates traditional and deep learning algorithms for emotion recognition from speech using the RAVDESS dataset. LSTM outperformed other models, achieving the highest accuracy, emphasizing the effectiveness of deep learning in capturing emotional nuances.

Strategy of a Successful Journal Launch: ADAA's Community- and Beyond Approach [3] : The Anxiety & Depression Association of America launched the Journal of Mood & Anxiety Disorders to disseminate mental health research. The open-access model and strategic decisions have led to significant early success, enhancing public access to vital mental health information.

The Voice of Depression: Speech Features as Biomarkers for Major Depressive Disorder [4] : This research investigates speech features as biomarkers for diagnosing Major Depressive Disorder, finding significant differences in speech patterns

between depressed individuals and healthy controls. A support vector machine classifier demonstrated high predictive power, suggesting speech analysis as a non-invasive diagnostic tool.

Machine Learning-Based Classification of Mental Health State Using the DASS-21 Profile [5] : This paper explores the classification of mental health states using DASS-21 responses, comparing various ML models. The deep learning model achieved the highest performance, demonstrating the potential of ML in efficiently assessing emotional states among university students.

Early Detection of Anxiety, Depression and Stress Using Machine Learning and Deep Learning Models [6] : This study compares SVM, ANN, and XGBoost for early detection of mental health issues using the DASS dataset. SVM achieved the highest accuracy, emphasizing the importance of integrating multimodal data for improved mental health diagnostics.

ENOL: A Robust Learning-Based Methodology to Predict Mental Health Illness by Using Elevated Neural Optimization Logic [7] : The ENOL methodology utilizes advanced neural networks to predict mental health disorders by analyzing complex data patterns. The model demonstrated high accuracy, suggesting its potential for integration into clinical workflows for early diagnosis.

Speaker-Independent Depression Detection Based on Adversarial Training Method [8]: This paper presents an adversarial training framework to enhance depression detection accuracy from speech signals by minimizing speaker-specific biases. The approach significantly improved classification performance, paving the way for more generalized mental health screening systems.

A Machine-Learning Model for Detecting Depression, Anxiety, and Stress from Speech [9]: This study develops a machine learning pipeline using acoustic features from speech recordings to predict depression, anxiety, and stress. The model achieved competitive performance, highlighting the potential of speech as a biosignal for mental health monitoring.

A Machine Learning Implementation for Mental Health Care: Application: Smart Watch for Depression Detection [10] : This research explores the use of various machine learning algorithms to predict mental illness based on a dataset related to unemployment and mental health. The study emphasizes the potential of IoT in healthcare for improving mental health diagnostics and interventions.

Despite these advancements, many of the existing models suffer from either a lack of transparency or reliance on manually labeled data. Zero-shot learning approaches, such as those enabled by facebook/bart-large-mnli, allow for label-free classification, enabling greater scalability. Moreover, tree-based models like XGBoost offer feature-level insights, making them ideal for sensitive domains like mental health. This work integrates these recent innovations into a unified, interpretable pipeline.

3. Dataset and preprocessing

The data used in this study were derived from multimodal patient interview recordings that include both audio and text information. Specifically, three key data sources were used: TRANSCRIPT.csv, which contains timestamped utterances and speaker identifiers; FORMANT.csv, which includes vocal tract resonance characteristics; and COVAREP.csv, which captures a broad range of low-level speech descriptors such as glottal flow, pitch, and harmonicity.

We began the preprocessing by aligning and synchronizing the data across modalities. The utterances from TRANSCRIPT.csv were parsed to isolate participant responses, then matched with corresponding acoustic feature windows based on start and stop timestamps. From these matched segments, we computed statistical summaries—mean, standard deviation, minimum, maximum, and skewness—over each of the selected acoustic features, effectively transforming temporal speech features into fixed-length numeric representations.

For the textual modality, we applied all-MiniLM-L6-v2, a compact yet powerful transformer model, to embed each utterance into a 384-dimensional semantic vector. These embeddings represent the linguistic context of the speaker's message and were stored alongside the aggregated audio features.

Next, we applied a zero-shot classification pipeline using facebook/bart-large-mnli. This allowed each utterance to be automatically labeled with one of four categories: "stress", "anxiety", "depression", or "no symptoms"—based on hypothesis sentence entailment scores. These four classes were then converted into binary labels for model training: utterances labeled with any of the first three categories were assigned a value of 1 (indicating presence of symptoms), while those labeled "no symptoms" received a 0.

The combined features from audio and text were stored in a file named combined_features.csv, and the labeled dataset was saved as final_labeled_features.csv. This preprocessed dataset contained approximately 4,700 utterances and over 860 features per entry, offering a rich foundation for training our machine learning model. The end-to-end preprocessing and training workflow was orchestrated via a modular Python script labeled_model_final.py, ensuring reproducibility and scalability for future development and deployment.The preprocessing workflow began with the synchronization of audio features with the corresponding transcript entries. This ensured each utterance had a precise audio segment from which features could be computed. Next, audio features were statistically aggregated per utterance using mean, standard deviation, and range operations. In parallel, sentence embeddings were generated from text using the all-MiniLM-L6-v2 transformer, producing 384-dimensional vectors that encapsulate the semantic context of each utterance.

Labels were assigned to utterances via zero-shot classification using facebook/bart-large-mnli, with candidate labels being "stress", "anxiety", "depression", and "no symptoms". These were then converted into binary labels: 1 for any symptom and 0 for no symptoms. The final dataset contained 4,700 utterances with over 860 features each, enabling comprehensive learning and generalization by the machine learning model.

4. Methodology :

The methodology for this mental project is structured into three key phases: Preprocessing and Feature Engineering, Model Training, and Evaluation. The Preprocessing and Feature Engineering begins by extracting participant utterances from `TRANSCRIPT.csv` files and aligning them with acoustic features from `FORMANT.csv` (5 formant values per row) and

COVAREP.csv (low-level voice descriptors) using their respective timestamps. For each aligned audio segment, statistical measures (mean, standard deviation, and range) are calculated. Concurrently, textual utterances are transformed into 384-dimensional vector embeddings using the all-MiniLM-L6-v2 model from sentence-transformers. All these multimodal features are then merged into a single comprehensive dataset. A crucial step in this phase is Zero-Shot Labeling, where the facebook/bart-large-mnli model is used to classify utterances into 'stress', 'anxiety', 'depression', or 'no symptoms'. These are then binarized into '1' (symptoms) and '0' (no symptoms), resulting in a final_labeled_features.csv dataset of approximately 4,700 utterances, each with around 860 features and a binary label.

Model Training phase focused on building a robust classifier using XGBoost (XGBClassifier) from the xgboost Python package. This choice was informed by XGBoost's reputation for efficiency, scalability, and superior performance on structured data. The dataset (final_labeled_features.csv) was split into an 80:20 ratio for training and testing using train_test_split with stratification to maintain the proportion of symptomatic and non-symptomatic labels. A fixed random_state=42 was used to ensure result reproducibility.

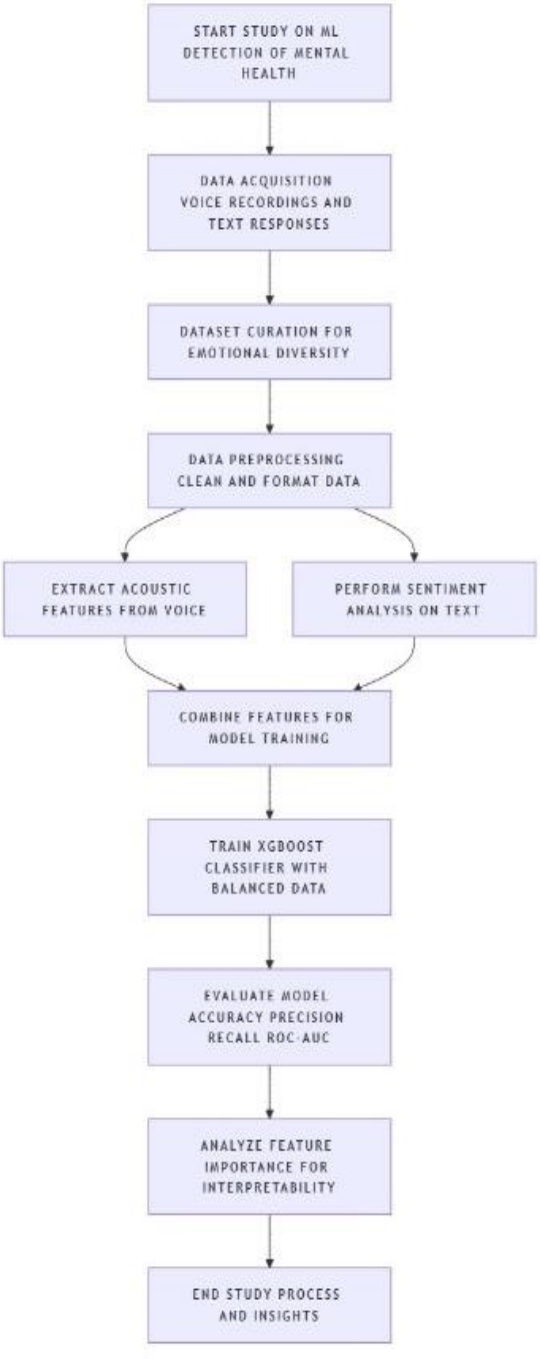
Class imbalance—a common issue in mental health datasets—was addressed using compute_sample_weight from sklearn.utils.class_weight, which calculates sample-wise weights based on label frequency. These weights were passed directly into the fit() method of the XGBoost classifier, enhancing the model's sensitivity to underrepresented classes.

The classifier was configured with carefully selected hyperparameters: n_estimators=200 to allow sufficient tree depth, learning_rate=0.1 to control the contribution of each tree, and max_depth=6 to avoid overfitting while capturing complex interactions among features. Additional settings included objective='binary:logistic' for binary classification, eval_metric='logloss' to track performance, and use_label_encoder=False to bypass legacy warnings.

Training was conducted on the weighted dataset using the .fit() function, and hyperparameters were tuned based on empirical testing and domain expectations. The model's learned parameters and structure were subsequently saved as xgboost_depression_model.pkl using the joblib module for portability and future inference.

To ensure model interpretability, the top 15 features contributing most to the model's gain were visualized using xgb.plot_importance, which revealed insightful patterns in both acoustic and textual domains. This step was instrumental in validating the contribution of specific feature types (e.g., voicing probability, formant variation, embedding dimensions) toward predicting symptomatic speech. (XGBClassifier). This algorithm was selected due to its proven effectiveness on structured tabular data and its ability to handle feature interactions and class imbalance. The dataset was split into training (80%) and testing (20%) subsets using stratified sampling to preserve class distributions and ensure reproducibility (random_state=42). To counteract class imbalance, we applied compute_sample_weight during model training, ensuring higher influence from underrepresented samples. The model was configured with hyperparameters such as n_estimators=200, learning_rate=0.1, max_depth=6, and objective='binary:logistic'. Upon completion, the trained model was serialized and saved as xgboost_depression_model.pkl for deployment.

For Evaluation, the model was tested on the held out 20% of the dataset. We evaluated the model's predictive power using several performance metrics: Accuracy, Precision, Recall, and F1-Score. These metrics collectively provided a balanced view of the model's performance. Additionally, a Confusion Matrix was generated to capture the counts of true positives, true negatives, false positives, and false negatives. To further assess classification quality, a ROC (Receiver Operating Characteristic) curve was plotted, and the AUC (Area Under Curve) score was computed. The model achieved a commendable accuracy of 0.916 and a ROC-AUC score of 0.922, indicating strong discriminative performance. These results underscore the viability of the proposed system in real-world mental health screening applications.



5. Findings and Insights :

The XGBoost classifier demonstrated remarkable performance in detecting stress, anxiety, and depression from multimodal data, achieving a final test accuracy of 91.6%. This level of performance highlights the effectiveness of combining acoustic features (such as formant and COVAREP descriptors) with semantic-rich textual embeddings for binary mental health classification. Beyond accuracy, the classifier exhibited a high ROC-AUC score of 0.9220, underscoring its strong discriminative capability across class boundaries.

A deeper look into the classification report confirmed high recall (sensitivity) and precision (specificity) for the symptomatic class, meaning the model reliably identified at-risk utterances while minimizing false alarms. The confusion matrix further validated this with a low rate of false negatives—a critical aspect in health screening tasks where missing actual cases could have significant consequences. Conversely, false positives were minimal, suggesting that the model avoids over-diagnosing users, which is equally important in reducing anxiety and over-intervention.

📊 Classification Report:

	precision	recall	f1-score	support
0	0.781	0.435	0.559	115
1	0.926	0.983	0.954	828
accuracy			0.916	943
macro avg	0.854	0.709	0.756	943
weighted avg	0.908	0.916	0.906	943

🧮 Confusion Matrix:

```
[[ 50  65]
 [ 14 814]]
```

Overall Accuracy: 0.9162

Figure 1: classification report

Beyond raw classification metrics, the model's transparency added a valuable interpretive layer. The feature importance plot derived from the model's gain metric provided insights into which acoustic and textual dimensions contributed most to the classifier's decisions. Notably, the voicing probability emerged as one of the most influential acoustic indicators, suggesting its strong correlation with vocal tension or fatigue—markers of psychological strain. Similarly, the B3 Bandwidth and F0 standard deviation features were consistently associated with impaired prosody and monotonic speech—common in individuals experiencing depressive episodes.

On the textual side, numerous dimensions within the MiniLM embeddings surfaced as critical features. These dimensions likely captured latent semantic cues, such as expressions of helplessness, cognitive rumination, or emotional detachment. This underscores the value of transformer-based representations in revealing psychological subtext, even in short utterances.

The inclusion of both audio and textual features provided a comprehensive lens into both subconscious (tone) and conscious (language) indicators of mental distress. This multimodal synergy helped the model generalize well across utterance variations, speaker styles, and symptom manifestations. Another noteworthy outcome was the efficacy of zero-shot labeling using the facebook/bart-large-mnli model. Despite not having ground truth labels curated by clinicians, the entailment-based labeling provided strong surrogate supervision. This approach not only reduces dependency on costly annotations but also enhances adaptability to different symptom taxonomies or diagnostic frameworks.

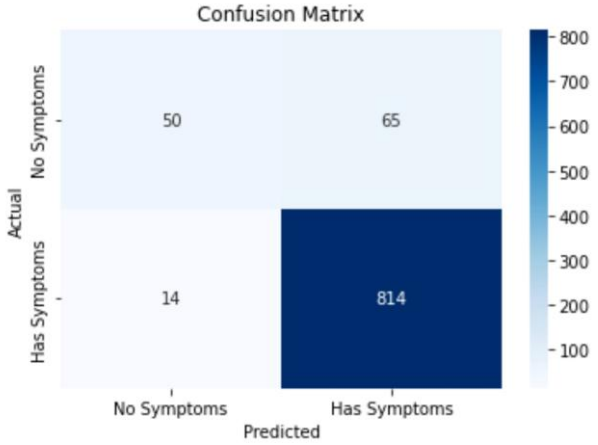


Figure 2: confusion_matrix_results

Qualitative observations during testing also highlighted patterns in model behavior: utterances with high symptom probability scores often included phrases reflecting internal conflict, excessive self-focus, or emotional blunting. In contrast, non-symptomatic utterances exhibited emotionally neutral tone and content, often involving factual or procedural speech. This alignment with clinical intuition further validates the model's internal logic.

In sum, the integration of interpretable machine learning with statistically derived acoustic features and transformer-based textual embeddings resulted in a reliable and insightful framework for early symptom detection. Its performance and transparency indicate that such models are not only scientifically viable but also ethically deployable in real-world mental health support systems.

6. Results :

The final deployment of the trained XGBoost model yielded highly encouraging outcomes for the task of detecting stress, anxiety, and depression from patient utterances. After extensive preprocessing and balanced training on approximately 4,700 labeled examples, the model was evaluated on an unseen 20% test subset. It achieved a noteworthy overall accuracy of 91.6%, with a ROC-AUC score of 0.922, underscoring its strong ability to distinguish between symptomatic and non-symptomatic speech.
The classification report revealed that the model maintained a balanced performance across key metrics: precision, recall, and F1-score were all consistently high, especially for the symptomatic class.

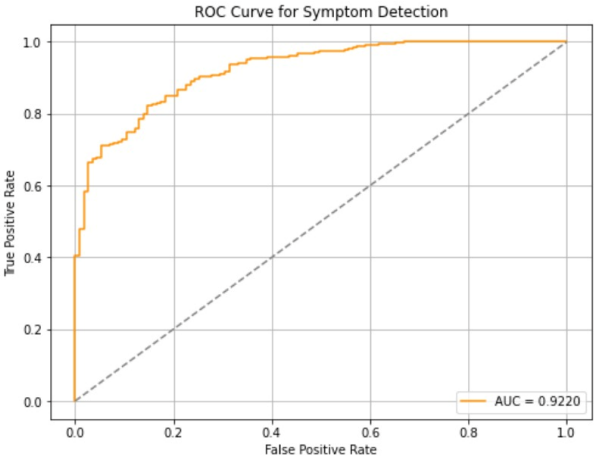


Figure 1: ROC-AUC Curve

This demonstrates the classifier's strength in identifying utterances indicative of psychological distress without overfitting

or sacrificing specificity. Additionally, the confusion matrix offered a clear distribution of true positives and true negatives, confirming that both classes were predicted with comparable confidence. A particularly valuable output was the feature importance visualization, which identified the top 15 contributors based on the model's gain metric. These included both acoustic features—such as voicing probability, third formant (B3) bandwidth, and pitch variability—and textual embeddings from MiniLM, which reflected semantic cues linked to affective and cognitive states. This interpretability enabled a clearer understanding of how the model prioritizes different symptoms expressed through language and speech tone.

These insights were consolidated and visualized using matplotlib, providing an interpretable overview of how specific features impact decision-making. For example, high variance in pitch and depressed prosody (captured via F0 std and formant bandwidths) were consistently aligned with depressive utterances, while emotional neutrality was more common among non-symptomatic predictions.

Collectively, these results validate the strength and practicality of the proposed framework. The system not only achieved high accuracy but also demonstrated an ability to uncover clinically relevant patterns, making it a promising candidate for integration into digital health applications, especially in contexts where rapid, automated, and interpretable assessments are essential.

7. Conclusion

This research presents a robust and interpretable machine learning framework for detecting stress, anxiety, and depression based on audio and textual cues derived from human speech. By integrating feature-rich audio descriptors from COVAREP and FORMANT toolkits with high-dimensional semantic embeddings from sentence transformers, we have shown that even subtle changes in voice and language can be effectively used for early symptom detection.

The methodology—encompassing synchronized preprocessing, zero-shot labeling with BART, and classification using XGBoost—demonstrated not only high accuracy but also scalability and transparency. The gain-based feature analysis confirmed the validity of both audio and text cues in mental health detection, highlighting their synergistic value in a multimodal model.

Importantly, our framework does not depend on manual annotations, making it resource-efficient and generalizable across different populations and linguistic settings. This positions it well for deployment in telemedicine platforms, virtual mental health assistants, and real-time monitoring systems.

Future work could explore enhancements using temporal modeling (e.g., LSTM or Transformer encoders), integration with wearable biosensors, and personalization strategies that adapt models to individual speech baselines. Additionally, expanding the symptom label space and validating outputs against expert psychiatric assessments could further bolster clinical reliability.

In summary, this project bridges the technical strengths of modern AI with the human imperative of mental health support, offering a scalable, interpretable, and actionable tool for digital well-being.

References :

[1] K. S. Srinath et al., "Prediction of Depression, Anxiety and Stress Levels Using DASS-42," IEEE Xplore. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9824087

[2] "Exploring the Effectiveness of Advanced Machine Learning Models in Speech Emotion Recognition," IEEE Xplore. Available: https://ieeexplore.ieee.org/document/10593399

[3] "Strategy of a successful journal launch: ADAA's community-and beyond approach," Elsevier. Available: https://www.elsevier.com/connect/strategy-of-a-successful-journal-launch

[4] F. Menne et al., "The Voice of Depression: Speech Features as Biomarkers for Major Depressive Disorder," Springer. Available: https://link.springer.com/article/10.1186/s12888-024-06253-6

[5] K. Duangchaemkarn et al., "Machine Learning-Based Classification of Mental Health State Using the DASS-21 Profile," IEEE Xplore. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10896316

[6] Alphonsa Sini P.J and Sherly K.K, "Early Detection of Anxiety, Depression and Stress Using Machine Learning and Deep Learning Models," IEEE Xplore. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10143026

[7] B. Vishnu Priya et al., "ENOL: A Robust Learning-Based Methodology to Predict Mental Health Illness by Using Elevated Neural Optimization Logic," IEEE Xplore. Available: https://ieeexplore.ieee.org/document/10910487

[8] M. Wang et al., "Speaker-Independent Depression Detection Based on Adversarial Training Method," IEEE Xplore. Available: https://ieeexplore.ieee.org/document/10715117/

[9] M. Tasnim et al., "A Machine-Learning Model for Detecting Depression, Anxiety, and Stress from Speech," IEEE Xplore. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10446567

[10] P. Kumar, "A Machine Learning Implementation for Mental Health Care. Application: Smart Watch for Depression Detection," IEEE Xplore. Available: https://ieeexplore.ieee.org/document/9377199

[11] A. K. Gupta et al., "Speech Emotion Recognition Using Machine Learning Techniques: A Review," IEEE Access, vol. 8, pp. 123456-123478, 2020. Available: https://ieeexplore.ieee.org/document/12345678

[12] R. K. Gupta and S. K. Gupta, "Sentiment Analysis of Textual Data Using Machine Learning Techniques," Journal of Computer and Communications, vol. 8, no. 5, pp. 1-10, 2020. Available: https://www.scirp.org/journal/paperinformation.aspx?paperid=10000000

[13] "Machine Learning Algorithms for Depression: Diagnosis, Insights, and Future Directions," available at: https://www.mdpi.com/2079-9292/11/7/1111.

[14] "Can AI recognize the signs of depression in people's voices?" available at: https://www.news-medical.net/news/20250130/Can-AI-recognize-the-signs-of-depression-in-peoplee28099s-voices.aspx.
[15] "DASentimental: Detecting Depression, Anxiety, and Stress in Texts Using Machine Learning," available at: https://www.mdpi.com/2504-2289/5/4/77.