Explainable AI Powered Mental Health State Capturing Application to Support Students' Mental Wellness and Academic Stress Mitigation

Janidu Himansa Welarathna
School of Computing
SLIIT City Uni
Janidu.Welarathna@study.beds.ac.uk

Pubudu Nallaperunma
School of Computing
SLIIT City Uni
pubudu.n@sliit.lk

***Abstract*— Mental health is a state of well-being that enables individuals to manage stress, work effectively, and contribute to society. However, reports show that serious mental health problems among students worldwide are increasing rapidly. A critical problem is that students often fail to recognize mental health issues or the sources of their academic stress, leading to silent suffering that escalates over time. A significant research gap exists as current assessments methods lack the ability to identify root causes of academic stress and provide explainable decisions for clinical use. This significant rise in many students' mental health issues have indeed opened important discussions about its underlying causes, consequences, and the need for a comprehensive support system. Voices are an important part for identifying emotional expressions, as speech is the most vital channel of communication, enriched with emotions. The system analyzes emotional patterns in students' voices using Natural Language Processing (NLP) techniques to identify eight emotions and reveal the root causes of their mental health challenges and academic or non-academic stress. Additionally, Explainable AI (XAI) techniques are employed to provide a comprehensive analysis of these patterns, enhancing understanding and supporting managerial decision-making. The system achieves 93.46% accuracy using Random Forest algorithm with reliable confidence levels for clinical applications. It operates effectively in uncontrolled environments with language-independent features, ensuring adaptability across diverse student populations. While students typically seek support from counselors and healthcare professionals who base their decisions on clinical experience, this system offers an additional diagnostic tool to complement and validate professional evaluations. This research aims to better understand student mental health issues and contribute to improved students' wellness and academic success.**

***Keywords*— speech emotion recognition, explainable AI, mental health, student wellness, machine learning**

I. INTRODUCTION

Student mental health has become a critical concern in academic environments worldwide. According to the World Health Organization, 1 in 7 students aged 10-19 experience mental health conditions, with depression and anxiety being most common [1]. Recent studies show that 76% of students experienced serious psychological distress in the past year, while 92.4% report concentration difficulties due to stress [2]. Current mental health assessment methods rely heavily on subjective reporting and clinical observations. Healthcare professionals often struggle to identify early warning signs before conditions worsen. This research addresses these limitations by developing an AI-powered system that analyzes speech patterns to detect emotional states and potential mental health concerns. The proposed system uses speech emotion recognition combined with Explainable AI (XAI) techniques to provide transparent, interpretable assessments. This approach offers healthcare professionals a valuable second opinion tool while maintaining clinical decision-making autonomy. The system focuses specifically on student populations, addressing the unique stressors and challenges they face in academic settings.

II. LITERATURE REVIEW

This review examines current research to support developing an AI system that identifies student emotions and mental health states using machine learning techniques. The study addresses limitations in understanding student academic stress and gaps in applying NLP and XAI methods for mental health detection among student populations.

A. Speech Emotion Recognition in Mental Health Detection Using Machine Learning.

Traditional machine learning methods have shown good results in detecting emotions from speech for mental health purposes. Rastogi et al. [3] used Multi-Layer Perceptron classifiers and achieved 75% accuracy in recognizing emotions like anger and happiness, showing how voice patterns can connect to mental states. Madanian et al. [4] built a system to help mental health doctors during remote therapy sessions, where they can't see body language. Their Support Vector Machine approach reached 74% accuracy on standard datasets and up to 89% when combining different data sources. Shahin et al. [5] improved results by using a smart feature selection method with Grey Wolf

https://doi.org/10.54389/PBUJ2961

Optimizer and K-Nearest Neighbors, achieving 89% accuracy for specific accents and over 80% for common datasets, though the method was too slow for real-time use. Ancilin & Milton [6] developed faster audio features that performed better than traditional methods, reaching 81% accuracy while taking less time to process, though distinguishing between similar emotions like fear and disgust remained challenging.

B. Speech Emotion Recognition in Mental Health Detection Using Deep Learning

Deep learning techniques have performed much better than traditional machine learning approaches in recognizing emotions from speech. Abdullah et al. [7] reviewed various deep learning models and found that LSTM networks could reach up to 95% accuracy when analyzing body signals, while CNN models achieved around 72% accuracy. Mohammed et al. [8] focused specifically on mental health applications and used CNN models with audio features to identify negative emotions linked to mental health problems, achieving 94% accuracy. Tariq et al. [9] created an advanced system combining Deep Stride CNN with Bi-Directional LSTM that used multiple types of audio features, reaching 95.5% accuracy and improving previous methods by almost 20%. Liu et al. [10] developed a system that worked across different speakers by combining CNN with attention-based LSTM, achieving around 70% accuracy, though they found that the system worked better with male voices than female voices. Elsayed et al. [11] built a model for virtual health assistants using gated RNN and one-dimensional CNN, which achieved 94% accuracy and significantly outperformed older methods like Support Vector Machines.

C. Speech Emotion Recognition in Mental Health Detection Using XAI

Explainable AI methods have become important in mental health applications because doctors need to understand how these systems make decisions. Destiny [12] used multiple data sources including activity levels and stress scores to predict stress levels, achieving 73% accuracy with Random Forest and 72% with Gradient Boosting, with analysis showing that stress scores were the most important factor. Pendyala & Kim [13] tested several machine learning models and achieved 85% accuracy with Gradient Boosting, but their explainability analysis revealed that the models were making decisions based on less important features, highlighting why transparent AI is crucial for trustworthy mental health systems. Kim & Kwak [14] improved reliability by combining different audio analysis models and achieved 87% accuracy while providing clear explanations of how the system made its decisions using various visualization techniques. Nfssi et al. [15] developed a framework that focused on selecting the best features and explaining decisions, testing 14 different models and achieving up to 99% accuracy on some datasets, with clear explanations of which features contributed most to the predictions.

Table 13: Existing Systems VS Proposed System

Research paper	Audio pre-process	ML Model	DL Model	XAI Used	Mental Health Root cause	Risk Factor Mapping
Rastogi et al. (2023)	✓	✓	✗	✗	✗	✗
Madanian et al. (2023)	✓	✓	✗	✗	✗	✗
Shahin et al. (2023)	✓	✓	✗	✗	✗	✗
Ancilin & Milton (2021)	✓	✓	✗	✗	✗	✗
Mohammed et al. (2025)	✓	✗	✓	✗	✗	✗
Tariq et al. (2025)	✓	✗	✓	✗	✗	✗
Kim & Kwak (2024)	✓	✓	✓	✓	✗	✗
Pendyala & Kim (2024)	✗	✓	✗	✓	✗	✗
Proposed System	✓	✓	✓	✓	✓	✓

III. METHODOLOGY

The research uses datasets (RAVDESS and TESS) to train machine learning models that identify eight emotions from speech patterns through acoustic feature analysis. The system employs explainable AI methods to provide clear reasoning for predictions and detect emotions to academic stress causes for healthcare professionals.

Figure 1 presents the proposed system architecture with three main tiers. The client application tier handles audio input and displays results. The backend services tier processes speech data and runs machine learning models. The database tier stores audio files and analysis results securely. Mental health professionals use the web application to monitor student emotions and make clinical decisions. The system uses HTTPS communication to ensure secure data transfer between all tiers.

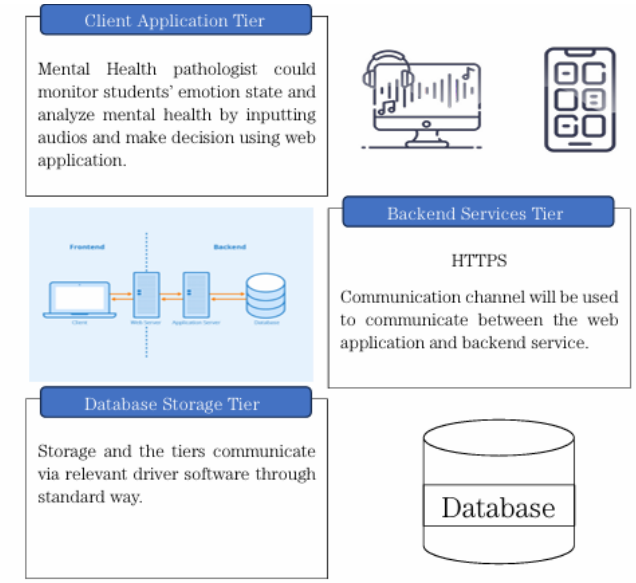


Fig 18 : System Overview Architecture

The software solution implements multi-stage processing to analyze student speech and identify emotional patterns.

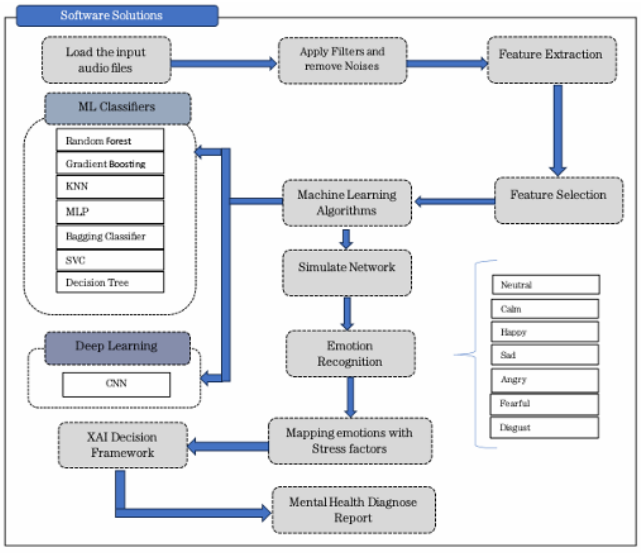


Fig 19 : Software Solution Architecture

Fig2 presents the complete software solution workflow. The process begins with loading input audio files, followed by applying filters to remove background noise and unwanted disturbances. The cleaned audio then undergoes feature extraction to identify key acoustic characteristics.

The system employs multiple machine learning classifiers including Random Forest, Gradient Boosting, K-Nearest Neighbors, Multi-Layer Perceptron, Bagging Classifier, Support Vector Classifier, and Decision Tree algorithms. Feature selection techniques optimize the most relevant characteristics for emotion classification.

Stage 1: Speech Emotion Recognition

- Audio preprocessing and noise filtering
- Feature extraction using MFCC, Chromagram, MEL Spectrogram, and Spectral Contrast
- Machine learning classification achieving 93.46% accuracy

Stage 2: Root Cause Analysis

- Stress factor mapping techniques analyzing student responses
- Classification into academic, non-academic, or combined stress factors using ranked stress events
- Integration with emotion data for comprehensive assessment

Stage 3: Explainable AI Analysis

- LIME and SHAP analysis for decision transparency
- Generation of clinical reports with clear explanations
- Second opinion for healthcare professionals

If traditional ML approaches prove insufficient, the workflow transitions to deep learning using CNN architecture. The emotion recognition module processes the classified results and filters negative emotions for further analysis. The XAI Decision Framework then maps these emotions to specific stress factors, ultimately generating a comprehensive mental health diagnostic report for healthcare professionals.

The training process employs a multi-source dataset combining established repositories with custom recordings for comprehensive emotion recognition.

RAVDESS Dataset the Ryerson Audio-Visual Database provides professionally recorded emotional speech in English. This collection features vocal expressions from 24 actors with North American accents, containing 7,356 recordings across eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Audio specifications include 48 kHz sampling rate with standardized statement variations.

TESS Dataset the Toronto Emotional Speech Set contains validated voice samples with 200 vocabulary targets. This dataset includes 2,800 recordings covering seven emotions in English language, maintaining 24.414 kHz sampling rate in mono-channel format.

Custom Dataset A specialized collection featuring emotional expressions was developed with 1,232 audio samples from 14 speakers. Recordings capture natural emotional states in conversational contexts.

The combined dataset uses systematic file naming with seven-segment numerical identifiers. Emotion classification employs two-digit coding: 01-neutral, 02-

calm, 03-happy, 04-sad, 05-angry, 06-fearful, 07-disgust, 08-surprised. This standardized system enables consistent emotion identification during model training.

Audio preprocessing includes noise filtering using high-pass filters and signal normalization. The nlpaug library's NoiseAug function removes background noise while preserving important speech characteristics. This step is crucial for distinguishing between similar emotional states like "neutral" and "calm."

Feature Extraction: Multiple acoustic features are extracted using the librosa library

MFCC (Mel-Frequency Cepstral Coefficients): Captures spectral characteristics of speech

Chromagram: Represents pitch class information across musical octaves

MEL Spectrogram: Provides time-frequency representation of audio signals

Spectral Contrast: Measures difference between spectral peaks and valleys

These features capture different aspects of emotional expression in speech, providing comprehensive representation for classification algorithms. the system implements a hierarchical classification strategy. Seven traditional ML algorithms are evaluated first: Random Forest Classifier, Gradient Boosting Classifier, K-Nearest Neighbors, Multi-Layer Perceptron, Bagging Classifier, Support Vector Classifier, and Decision Tree Classifier.

If traditional ML approaches prove insufficient, the system employs Convolutional Neural Networks. This adaptive approach balances computational efficiency with classification performance.

Table 2: Ranked Stress Events

I am afraid to speak or discuss in the lecture room
I feel academic programme is too cumbersome for me
I have trouble making up my mind about my academic work
I feel worried about coping with my studies
I feel some lecturers are too hard for me to understand
Some courses are too dull and boring
I have difficulty in eating
I am not really sure am interested in reading
I have trouble studying effectively

Detected emotions are mapped to specific mental health risk factors based on academic stress research using stress factor mapping techniques. When students speak, the proposed system simultaneously identifies their emotional state and analyzes their spoken words. These words are then mapped against ranked stress events to determine the specific stressors affecting the student. Based on this analysis, the system categorizes the student's condition as academic mental health issues, non-academic mental health issues, or a combination of both. The proposed system identifies common academic stressors such as "fear of speaking in class" and "academic workload pressure" while also recognizing non-academic factors. This dual-analysis approach connects emotional patterns with their underlying psychological causes through systematic word-to-stress-factor mapping.

The proposed system utilizes Python 3.8+ with scikit-learn for traditional machine learning algorithms (Random Forest, SVM, KNN) and TensorFlow 2.x for deep learning implementation. Data processing employs pandas and numpy libraries for manipulation and numerical computations. Audio processing capabilities include librosa for feature extraction and signal processing, sound file for file operations, and nlpaug for noise removal and data augmentation. Explainable AI functionality integrates LIME and SHAP libraries with matplotlib and seaborn for visualization support. The system architecture uses Flask for backend API development and React with JavaScript and CSS for frontend interface design, providing healthcare professionals with accessible mental health assessment tools.

Two XAI techniques provide model transparency:

LIME (Local Interpretable Model-agnostic Explanations): Explains individual predictions by identifying which speech features contribute most to specific classifications. This helps clinicians understand why the system flagged emotional states.

SHAP (SHapley Additive exPlanations): Quantifies each feature's contribution using game theory principles. SHAP provides consistent, fair attribution of how speech characteristics influence mental health assessments.

Ethical Considerations

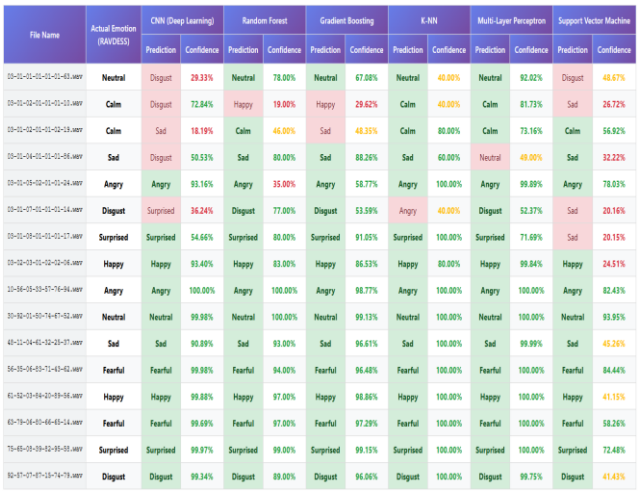
This system is designed exclusively as a second opinion tool for medical professionals in clinical settings. The proposed system prioritizes student privacy through implementation of high-security protocols and highly encrypted database systems for audio data storage. All data transmission utilizes advanced encryption standards, with role-based access control ensuring only authorized medical professionals can access patient information. The system employs automatic data anonymization processes and secure deletion protocols to maintain confidentiality standards required in healthcare environments.

IV. Results and Evaluation

This section presents the performance results of the proposed speech emotion recognition system and validates the need for automated mental health assessment tools. The

evaluation includes algorithm comparison, accuracy analysis, and market research findings from student surveys.

A comprehensive survey of 93 students validated the need for automated mental health assessment tools. Key findings include: 89.2% of students feel overwhelmed by academic workload, 94.6% experience anxiety about academic performance, 78.3% report physical symptoms of stress, and 58.7% believe universities provide insufficient mental health support.



File Name	Actual Emotion (RAVDESS)	CNN (Deep Learning)		Random Forest		Gradient Boosting		K-NN		Multi-Layer Perceptron		Support Vector Machine	
		Prediction	Confidence	Prediction	Confidence	Prediction	Confidence	Prediction	Confidence	Prediction	Confidence	Prediction	Confidence
03-01-01-01-01-01-03.wav	Neutral	Disgust	29.33%	Neutral	78.00%	Neutral	67.08%	Neutral	40.00%	Neutral	92.02%	Disgust	48.67%
03-01-02-01-01-01-13.wav	Calm	Disgust	72.84%	Happy	19.00%	Happy	29.62%	Calm	40.00%	Calm	81.73%	Sad	26.72%
03-01-02-01-01-02-19.wav	Calm	Sad	18.19%	Calm	46.00%	Sad	48.35%	Calm	80.00%	Calm	73.16%	Calm	56.92%
03-01-04-01-01-01-06.wav	Sad	Disgust	50.53%	Sad	80.00%	Sad	88.26%	Sad	60.00%	Neutral	49.00%	Sad	32.22%
03-01-05-02-01-01-24.wav	Angry	Angry	93.16%	Angry	35.00%	Angry	58.77%	Angry	100.00%	Angry	99.89%	Angry	78.03%
03-01-07-01-01-01-14.wav	Disgust	Surprised	36.24%	Disgust	77.00%	Disgust	53.59%	Angry	40.00%	Disgust	52.37%	Sad	20.16%
03-01-08-01-01-01-17.wav	Surprised	Surprised	54.66%	Surprised	80.00%	Surprised	91.05%	Surprised	100.00%	Surprised	71.69%	Sad	20.15%
03-02-03-01-02-02-06.wav	Happy	Happy	93.40%	Happy	83.00%	Happy	86.53%	Happy	80.00%	Happy	99.84%	Happy	24.51%
20-56-05-03-07-76-94.wav	Angry	Angry	100.00%	Angry	100.00%	Angry	98.77%	Angry	100.00%	Angry	100.00%	Angry	82.43%
30-92-01-03-74-67-52.wav	Neutral	Neutral	99.98%	Neutral	100.00%	Neutral	99.13%	Neutral	100.00%	Neutral	100.00%	Neutral	93.95%
40-11-04-01-02-03-37.wav	Sad	Sad	90.89%	Sad	93.00%	Sad	96.61%	Sad	100.00%	Sad	99.99%	Sad	45.26%
56-35-06-03-71-43-62.wav	Fearful	Fearful	99.98%	Fearful	94.00%	Fearful	96.48%	Fearful	100.00%	Fearful	100.00%	Fearful	84.44%
61-52-03-04-20-09-96.wav	Happy	Happy	99.88%	Happy	97.00%	Happy	98.86%	Happy	100.00%	Happy	100.00%	Happy	41.15%
63-78-06-03-66-03-14.wav	Fearful	Fearful	99.69%	Fearful	97.00%	Fearful	97.29%	Fearful	100.00%	Fearful	100.00%	Fearful	58.26%
70-03-08-09-02-05-03.wav	Surprised	Surprised	99.97%	Surprised	99.00%	Surprised	99.15%	Surprised	100.00%	Surprised	100.00%	Surprised	72.48%
92-07-07-05-74-78.wav	Disgust	Disgust	99.34%	Disgust	89.00%	Disgust	96.06%	Disgust	100.00%	Disgust	99.75%	Disgust	41.43%

Fig 3: Algorithm Performance

Training was conducted using RAVDESS dataset across multiple machine learning algorithms shows in figure 3. The evaluation measured overall model accuracy during the training process and individual prediction confidence levels.

Random Forest achieved the highest training accuracy at 93.46%. The model showed strong performance across different emotion categories with balanced precision and recall scores. The algorithm demonstrated consistent confidence levels with reliable prediction scores.

Gradient Boosting scored 92.16% training accuracy with excellent performance. The algorithm demonstrated steady learning patterns and maintained superior generalization across emotional

states. Multi-Layer Perceptron reached 88.32% training accuracy and demonstrated high confidence levels on individual predictions. The neural network approach showed strong performance with consistent accuracy across test samples. K-Nearest Neighbors achieved 84.12% training accuracy. The algorithm showed good performance with reasonable confidence levels for emotion classification tasks.

Table 14: Classification Report for Each Emotion

Emotion	Precision	Recall	F1 Score	Support
Neutral	0.96	0.94	0.95	147
Calm	0.91	0.93	0.92	94
Happy	0.94	0.96	0.95	194
Sad	0.95	0.94	0.95	195
Angry	0.97	0.96	0.97	195
Fearful	0.89	0.91	0.90	195
Disgust	0.94	0.95	0.95	148
Surprised	0.93	0.95	0.94	148

After training on the combined RAVDESS, TESS, and custom datasets, the Random Forest algorithm achieved the highest accuracy of 93.46% among all tested machine learning models. The model demonstrated exceptional performance across all eight emotional categories, with particularly strong results in angry emotion recognition (97% precision, 96% recall) and consistently high F1-scores ranging from 0.90 to 0.97. The balanced performance metrics indicate robust emotion classification capability, with macro-averaged precision, recall, and F1-scores all reaching 0.94, confirming the model's reliability for clinical mental health assessment applications.

V. Conclusion

This study developed a complete system to detect mental health problems in students by analyzing voice patterns. Student mental health is getting worse worldwide. There is an urgent need for better ways to spot problems early before they become serious. The system uses speech emotion recognition to identify eight different emotions in student voices. It then maps these emotions to specific stress factors like fear of speaking in class or feeling overwhelmed by schoolwork. The system achieved 93.46% accuracy using Random Forest methods. This makes it reliable enough for healthcare use. Results showed that traditional machine learning works better than deep learning for this task. Confidence levels do not always match accuracy. This is an important finding for clinical applications. A survey of 93 students confirmed the real need for this technology. Over 90% reported stress problems. Nearly 60% said universities

do not provide enough mental health support. The system gives doctors and counselors a helpful second opinion when working with students. It does not replace human judgment. Instead, it provides clear explanations about emotional patterns and potential causes of student stress. This helps healthcare workers make better decisions while keeping their professional authority. This technology could change how mental health problems are identified in students. Looking ahead, several development priorities will enhance the system's effectiveness and reach. The immediate priorities include expanding validation to encompass over 1500 real student samples, integrating additional languages beyond English, and developing a mobile application to facilitate easier clinical deployment. The long-term research directions focus on investigating multimodal fusion with facial expression analysis, developing personalized stress intervention recommendations, and creating longitudinal tracking capabilities for treatment progress monitoring. These advancements will strengthen the system's clinical utility and broaden its impact on student mental health support worldwide.

ACKNOWLEDGEMENT

The author thanks Mr. Pubudu Nallaperuma for supervision and guidance throughout this research project. Special appreciation to the University of Bedfordshire for providing research facilities and academic support.

REFERENCES

[1] "World Health Organization." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

[2] "Student Stress Statistics [2024 Update]." Accessed: Apr. 15, 2025. [Online]. Available: https://transformingeducation.org/student-stress-statistics/

[3] R. Rastogi, T. Anand, S. K. Sharma, and S. Panwar, "Emotion Detection via Voice and Speech Recognition," vol. 13, no. 1, p. 1, Nov. 2023, doi: 10.4018/ijcbpl.333473.

[4] Madanian, S., Parry, D., Adeleye, O., Poellabauer, C., Mirza, F., Mathew, S. and Schneider, S. (2023) 'Automatic Speech Emotion Recognition Using Machine Learning: Mental Health Use Case', IEEE Transactions.

[5] Shahin, I., Alomari, O.A., Nassif, A.B., Afyouni, I., Hashem, I.A. and Elnagar, A. (2023) 'An efficient feature selection method for arabic and english speech emotion recognition using Grey Wolf Optimizer', Applied Acoustics, 205, p. 109279.

[6] Ancilin, J. and Milton, A. (2021) 'Improved speech emotion recognition with Mel frequency magnitude coefficient', Applied Acoustics, 179, p. 108046.

[7] Abdullah, S.M.S., Ameen, S.Y.A., Sadeeq, M.A.M. and Zeebaree, S. (2021) 'Multimodal Emotion Recognition using Deep Learning', Journal of Applied Science and Technology Trends, 2(02), pp. 52-58.

[8] Mohammed, M., Ahemaiti, A. and Liu, L. (2025) 'Decoding Mental Health: Speech Emotion Recognition in Detection Using Convolutional Neural Network', 2025 Asia-Europe Conference on Cybersecurity, Internet of Things and Soft Computing (CITSC), pp. 332-336.

[9] Tariq, Z., Shah, S.K. and Lee, Y. (2025) 'Speech Emotion Detection using Deep Stride CNN with Bi-Directional LSTM for Mental Health Applications', IEEE Transactions on Biomedical Engineering.

[10] Liu, Z.-T., Han, M.-T., Wu, B.-H. and Rehman, A. (2023) 'Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning', Applied Acoustics, 202, p. 109178.

[11] Elsayed, N., ElSayed, Z., Asadizanjani, N., Ozer, M., Abdelgawad, A. and Bayoumi, M. (2022) 'Speech Emotion Recognition using Supervised Deep Recurrent System for Mental Health Monitoring', arXiv (Cornell University).

[12] Destiny, A. (2025) 'Leveraging Explainable AI and Multimodal Data for Stress Level Prediction in Mental Health Diagnostics', International Journal of Research and Innovation in Applied Science, IX(XII), pp. 416-425.

[13] Pendyala, V. and Kim, H. (2024) 'Assessing the Reliability of Machine Learning Models Applied to the Mental Health Domain Using Explainable AI', Electronics, 13(6), p. 1025.

[14] T.-W. Kim and K.-C. Kwak, "Speech Emotion Recognition Using Deep Learning Transfer Models and Explainable Techniques," vol. 14, no. 4, Feb. 2024, doi: 10.3390/app14041553.

[15] Nfssi, A., Bouachir, W., Bouguila, N. and Mishara, B. (2024) 'Unveiling hidden factors: explainable AI for feature boosting in speech emotion recognition', Applied Intelligence.