



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Psychological Stress Detection from Voice using Deep Learning

Ramyakrishna Kadiyala^{1*}, B Madhav Rao², Vyuhita Gunupudi³, Chanikya Durga Vara Prasad G⁴

Assistant Professor, Dept. of CSE, Sir C R Reddy College of Engineering, Eluru, India¹

Professor, Dept. of CSE, Sir C R Reddy College of Engineering, Eluru, India²

B. Tech Student, Dept. of CSE, Sir C R Reddy College of Engineering, Eluru., India^{3,4}

ABSTRACT: The prevalent problem of mental distress in contemporary society evolves to a severe issue that impacts mental and emotional conditions of people. The continuous existence of stress involves negative influence to cognitive functions and decision-making abilities and full working performance. The identification of psychological stress at an early stage allows its measurements to reduce their impact at a time when it improves the methods of treating mental health. The study presents a multimodal system of deep learning that identifies psychological distress by examining speech and written words. The system is based on the learning of particular acoustic features of speech data by a neural network model, CNN + LSTM, and a transformer-based language model that learns about the development of language patterns in written text. The researchers relied on the RAVDESS dataset to evaluate the speech emotion recognition and the Dreddit dataset to test the ability to detect textual stress. It was observed that the text-based model has accuracy of 86.8 and speech-based model is 96.11 and ROC-AUC= 0.93. The results suggest that the automated stress detection systems prove to be more effective when both acoustic and linguistic data are employed.

KEYWORDS: Stress Detection, Multimodal Learning, Speech Emotion Recognition, Deep Audio Signal Processing

I. INTRODUCTION

The contemporary world experiences rising psychological stress because of two major factors: technological progress and changing social norms and increasing work requirements. The human body functions better under moderate stress because it serves as a performance boost, but excessive stress periods lead to negative impacts on mental health and emotional balance and physical health. Previous research has established that chronic stress leads to multiple health problems which include anxiety disorders and depression and sleep disturbances and cardiovascular diseases [4][13]. The healthcare and artificial intelligence fields require researchers to create effective psychological stress detection techniques which have become essential for their work [10]. The traditional stress detection methods use subjective assessment through questionnaires and clinical interviews, and they measure stress through heart rate variability and skin conductance and brain activity signals [13]. The techniques enable stress level assessment, but they need medical equipment and qualified professionals who must include clinicians and psychologists in the process. The methods fail to work as monitoring tools for complete environmental observation. Speech, as a behavioral signal, enables people to assess their emotional and psychological condition through a method that requires no physical contact and is easy to obtain [8][9]. Human speech transmits deep acoustic and linguistic elements which demonstrate how speakers feel psychologically and emotionally. The different vocal elements of pitch and intensity and speech speed and speech rhythm, and speech patterns create vocal indicators that have value.

II. RELATED WORK

The detection of psychological stress has emerged as a significant field of research because it has been vital in the health care monitoring, workplace productivity, and human-computer interaction systems. Speech signals have been predominantly examined as a powerful channel of recognition of emotional and psychological conditions with early studies mainly relying on traditional machine learning methods in conjunction with manually engineered features which are derived out of physiological responses, speech patterns and textual communications [11], [12]. The initial ones were to generate acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC), pitch and energy and use them on classical machine learning classifiers to classify the emotional states based on speech signals, such as



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Support Vector Machines (SVM), k-Nearest Neighbors (KNN) and Random Forests [14], [11]. Indicatively, the classifiers included Decision Trees and KNN which were applied to categorize emotional states on the basis of extracted acoustic features [11]. With the advancement of deep learning techniques, researchers have increasingly adopted neural network architectures for stress detection and emotion recognition. Deep learning models are capable of automatically learning hierarchical feature representations from raw or minimally processed data, reducing the dependence on manual feature engineering [5], [12]. Convolutional Neural Networks (CNNs) have been popularly applied to spectrogram representations of speech signals to enable models to learn time frequency patterns of an emotional state. A number of studies have confirmed that CNN-based models are better than traditional machine learning techniques in speech emotion recognition tasks [15]. The most recent studies have also been conducted in terms of the integration of state-of-the-art artificial intelligence and automated machine learning strategies in order to enhance the efficiency and accuracy of intelligent systems in data analysis and prediction tasks [22].

Besides CNNs, Recurrent Neural Networks (RNNs) and specifically the Long Short-Term Memory (LSTM) networks, have been shown to be useful in learning sequential speech signal dependencies. A study conducted in [2] demonstrates that LSTM-based designs could be used to successfully extract the temporal relationships in speech sequence. Moreover, Bidirectional LSTM (Bidirectional long short memory) models are more effective as they process speech sequences backward and forward.

In addition to speech-based analysis, other researchers have studied the Natural Language Processing (NLP) methods in order to identify psychological conditions based on the textual communication. Transformer-based language models, including BERT and RoBERTa, have recently become very successful in different NLP-related tasks, like sentiment analysis, emotion recognition, and mental health prediction [18], [19]. Multimodal methods in detection of psychological stress have been started to be examined. Multimodal systems comprise integration of information obtained through several channels like speech, text, facial expression or physiological measures to enhance the accuracy of prediction [3]. Multimodal frameworks have the ability to offer a more encompassing insight into the state of human emotions and psychological pressures by taking into consideration the acoustic and contextual linguistic information.

Within this paper, we present a multimodal deep learning model of detecting psychological stress based on speech signals and contextual language information based on text signals. The proposed system will enhance the precision and strength of the automated stress detection systems by incorporating all the mentioned modalities.

Table 1 Literature survey of some of the existing work.

S. No	Author	Description	Methodology	Results
1	Cohen et al. [1]	Study of psychological stress and its impact on human health and disease.	Psychological and physiological stress analysis.	Demonstrated the relationship between stress and various health conditions.
2	Healey & Picard [2]	Detection of stress during driving tasks using physiological signals.	Physiological sensors and machine learning classification.	Successfully detected stress levels in real-world driving environments.
3	Zeng et al. [3]	Survey of affect recognition techniques using audio, visual, and multimodal signals.	Multimodal emotion recognition approaches.	Demonstrated improved emotion recognition using multimodal data.
4	El Ayadi et al. [4]	Comprehensive survey on speech emotion recognition techniques.	Feature extraction from speech signals and machine learning classifiers.	Highlighted the importance of acoustic features such as MFCC and pitch.
5	Hochreiter & Schmidhuber [5]	Introduction of Long Short-Term Memory (LSTM) networks for sequential data modeling.	Recurrent neural networks with memory cells.	Enabled effective modeling of temporal dependencies in speech signals.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

6	Krizhevsky et al. [6]	Deep convolutional neural networks applied to large-scale pattern recognition.	Convolutional Neural Networks (CNN).	Achieved state-of-the-art results in deep learning-based classification tasks.
7	Fayek et al. [7]	Evaluation of deep learning architectures for speech emotion recognition.	CNN-based deep learning models using speech features.	Achieved improved performance compared to traditional machine learning models.
8	Livingstone & Russo [8]	Development of the RAVDESS dataset for emotional speech analysis.	Audio-visual emotional speech dataset creation.	Provided a benchmark dataset widely used for speech emotion research.
9	Busso et al. [9]	Introduction of the IEMOCAP dataset for emotion recognition research.	Multimodal emotion dataset including audio, video, and motion capture.	Enabled research in multimodal emotion recognition.
10	De Choudhury et al. [10]	Detection of mental health conditions using socaposts.	Natural Language Processing on social media data.	Demonstrated potential of textual analysis for mental health prediction.
11	Devlin et al. [11]	Introduction of BERT for contextual language representation.	Transformer-based deep learning model for NLP.	Achieved significant improvements in NLP tasks including sentiment analysis.
12	Liu et al. [12]	Development of RoBERTa, an improved version of BERT.	Transformer-based language model with optimized training.	Demonstrated improved performance in text classification tasks.
13	McEwen [13]	Study of stress mediators and their biological effects.	Psychological and medical stress analysis.	Identified physiological mechanisms of stress.
14	Schuller et al. [14]	Speech emotion recognition using acoustic feature analysis.	Machine learning classifiers using speech features.	Demonstrated effectiveness of acoustic features for emotion recognition.
15	Eyben et al. [15]	Development of openSMILE toolkit for audio feature extraction.	Acoustic feature extraction framework for speech processing.	Widely used toolkit for speech emotion recognition research.
16	O'Shaughnessy [16]	Analysis of human speech communication and acoustic processing.	Speech signal processing techniques.	Provided foundational methods for speech analysis.
17	Han et al. [17]	Speech emotion recognition using deep neural networks.	DNN and feature extraction methods.	Improved classification accuracy for emotional speech.
18	Sarkar et al. [18]	Review of deep learning techniques for speech emotion recognition.	CNN, RNN, and hybrid deep learning models.	Demonstrated superiority of deep learning approaches over traditional methods.
19	Turcan & McKeown [19]	Introduction of Dreddit dataset for stress detection in social media text.	NLP-based stress detection using social media posts.	Provided benchmark dataset for textual stress detection.
20	Radford et al. [20]	Development of Whisper speech recognition model.	Large-scale deep learning speech recognition system.	Achieved robust speech transcription performance.
21	Bhandari [21]	Generation of log-Mel spectrogram features for audio analysis.	Signal processing using Librosa library.	Demonstrated effective feature extraction for speech analysis.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. PROPOSED ALGORITHM

In this section, the proposed multimodal deep learning model of detecting psychological stress based on speech and text is provided. The system is more effective in detecting stress as it employs more than one modalities to extract various information based on both the speech signals and textual communication. The general architecture comprises a few elements that are namely data acquisition, preprocessing, feature extraction, training of deep learning model, multimodal fusion, and final stress classification.

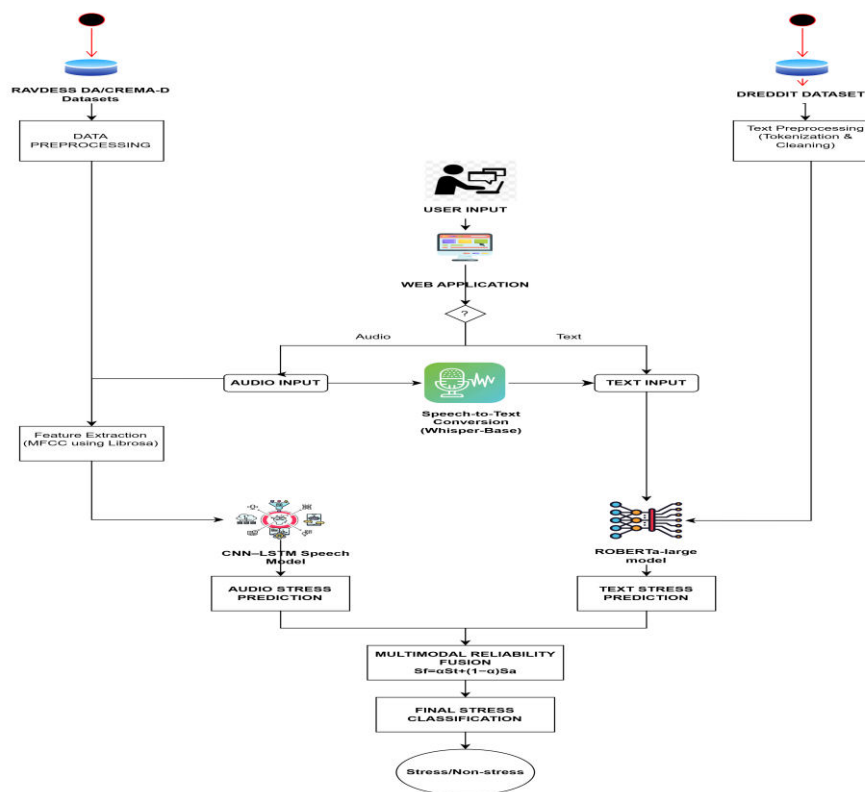


Fig-1. Architecture diagram

A. Dataset Description

The proposed stress detection framework uses various speech-based datasets and textual datasets for the training and testing of the proposed stress detection system. The speech-based datasets contain the acoustic features of emotional speech, while the textual datasets contain the linguistic features of emotional speech.

1. Speech Datasets

The proposed stress detecting system based on speech also utilizes the several available emotional speech databases publicly in training and testing of the proposed stress detection system.

- RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)
- CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset)

Table 2. Speech Datasets Used for Stress Detection

Dataset	Number of Samples	Speakers	Emotional Classes	Format
RAVDESS	1440	24	8 emotions	WAV
CREMA-D	7442	91	6 emotions	WAV



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

These datasets provide a diverse collection of emotional speech recordings that help the model learn generalizable acoustic patterns related to stress and emotional states.

2. Text Dataset

The Dreddit dataset is used to analyze the linguistic features that describe the concept of psychological stress. The dataset consists of social media content that users post on Reddit about their stressful experiences in their lives. The dataset contains the following two types of labeled data:

- Stress
- Non-Stress

Table 3. Text Dataset Used for Stress Detection

Dataset	Number of Posts	Labels	Source
Dreddit	~3500	Stress / non-stress	Reddit

B. Data Preprocessing

Since there might be some noise, inconsistency, and/or irrelevant information in raw speech and text data, data preprocessing is conducted. This is done to refine and improve the quality of data before it is fed into the deep learning models.

Speech Preprocessing

There are several speech-processing methods that involve signal processing techniques. These include:

- Noise reduction: This is conducted to remove some noise that might have interfered with the speech signals.
- Normalization: This is conducted to normalize speech signals.
- Silence removal: This is conducted to remove some silent moments in the speech signals.
- Resampling: This is conducted to ensure that all speech signals have the same sample frequency.

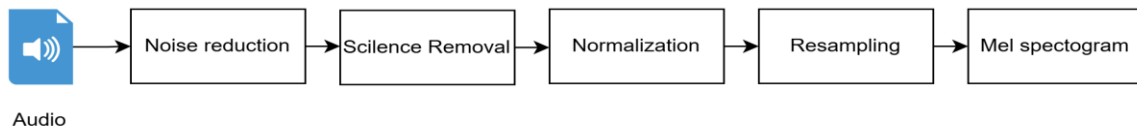


Fig-2: Audio Preprocessing and Feature Extraction Pipeline

Text Preprocessing

Text preprocessing is used for preparing the input data for the language model. The steps for text preprocessing are as follows:

- Tokenization of sentences
- Lower case conversion
- Removal of punctuation and unwanted symbols
- Text normalization

C. Audio Feature Extraction

The extraction of acoustic features is performed with the help of the Librosa library that is usually employed in machine learning models when working with audio signals.

The speech features that were important include:

- Mel Frequency Cepstral Coefficients (MFCC): This is a spectral representation technique that characterizes the characteristics of a melody (or speech) along the mel frequency.
- Mel Spectrogram
- Pitch and Energy Features



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The features obtained are relevant to the deep neural network to be applied in the stress recognition based on speech because the MFCC features can be used to depict the values of the speech signal in a way that is comparable to the human auditory system.

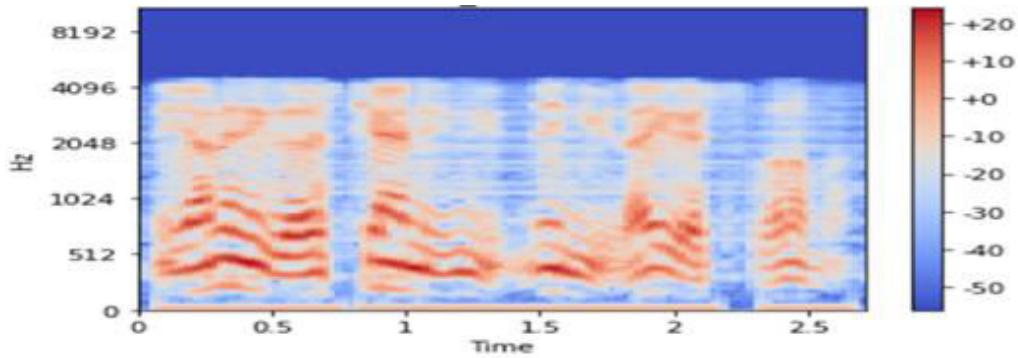


Fig-3. Example Mel-Spectrogram Representation of Speech Signal

D. Speech-to-Text Conversion

Whisper-Base, a state-of-the-art model of natural language processing, is also used to break down the speech recording into a written text in the proposed framework, which is analyzed through natural language processing. This operation enables the system to record the acoustic signs as well as the linguistic signs of speech signals.

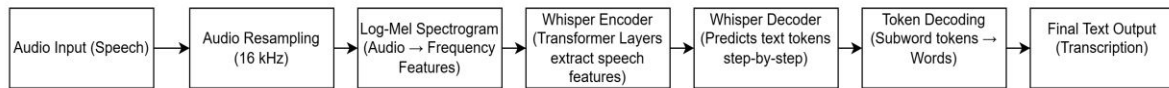


Fig-4: Speech-to-Text Transcription Process using Whisper Model

E. Deep Learning Model Architecture

The proposed system utilizes two deep learning models for processing the input audio and text data.

1. Audio Stress Detection Model (CNN-LSTM)

The suggested system employs the CNN-LSTM architecture to the audio stress detector module. Such an architecture can handle spatial as well as temporal characteristics of the speech signal. The local features of the MFCC feature map can be processed using the convolutional neural network layers. The sequential features of the speech signal can be taken care of by the LSTM neural network layers.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

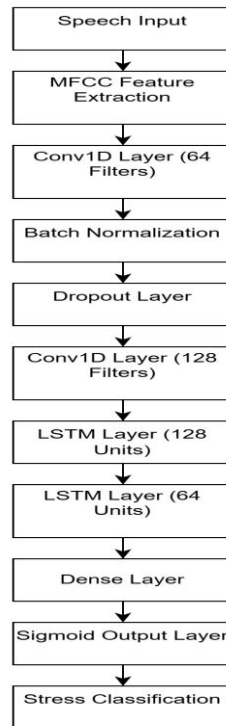


Fig-5: CNN-LSTM Architecture for Speech Stress Detection

2. Text Stress Detection Model

RoBERTa-Large model works with the text data and is a transformer-based language model that is specialized in contextual language understanding. The model uses a self-attention mechanism to bring out the capacity to understand associations among words in sentences. This allows it to do word embeddings that capture linguistic patterns which are associated with stressful or non-stressful conditions. The word embeddings are then processed in classification layers that determine whether the text assumes stress or non-stressful states.

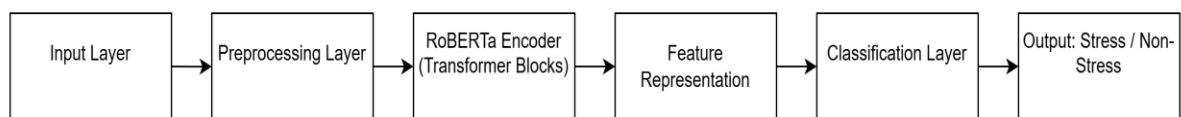


Fig-6: Architecture of the RoBERTa-Large Model for Text Stress Detection

F. Reliability Fusion Mechanism

After obtaining results from both models, the system fuses these results together through a reliability fusion approach. The fusion approach combines the results from both models and uses them to compute a final result that reflects stress conditions. This enables the system to leverage information from two modalities, thus improving the overall accuracy of the results.

$$S_f = \alpha S_t + (1 - \alpha)S_a$$

Where:

S_f – Final fused stress score

S_a – Speech model prediction

S_t – Text model prediction

α – Fusion weight



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. RESULTS AND DISCUSSION

This part describes the outcomes of the experiment conducted to assess the multimodal framework suggested to detect psychological stress. The effectiveness of the speech-based model, text-based model, and hybrid multimodal model is evaluated with a number of classification measures such as accuracy, precision, recall, F1-score, and ROC-AUC. These metrics will give a fine-grained assessment of the capability of the proposed system to determine stress-associated patterns using both speech indicators and written text.

A. Results of audio-based stress detection.

A stress detecting model based on speech is proposed, which is a CNNLSTM model, which is trained on acoustic signals that are extracted on speech. The most significant input representation of speech signals is Mel-Frequency Cepstral Coefficients (MFCC). The convolutional layers are the ones that learn local spectral statistics in the speech signal and the LSTM layers learn the temporal correlations in the sequence of audio frames. This conglomeration allows the model to acquire spectral and time-related attributes relating to emotional stress. The speech-based model performance assessment is based on a number of classification measurements such as accuracy, precision, recall, and F1-score.

Table 4. Performance of Speech Stress Detection Model

Model	Dataset	Accuracy	Precision	Recall	F1-Score
CNN-LSTM Speech Model	RAVDESS	0.9611	0.9172	0.8636	0.8896

These findings indicate that the acoustic features are effective in depicting emotional states. Diverse information about stress and emotional variations can be abundantly obtained in speech signals and well represented with deep learning architectures

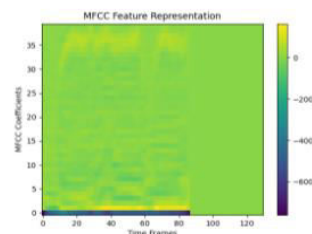


Fig-7. MFCC Feature Representation of Speech Signal



Fig-8. Audio Stress Detection Model Confusion Matrix.

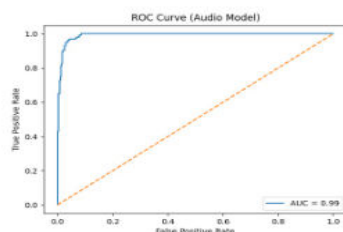


Fig-9. ROC Curve for CNN-LSTM Audio Stress Detection Model

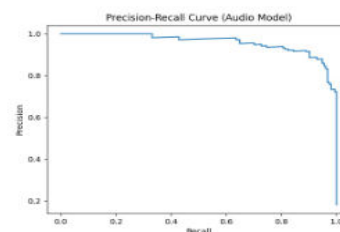


Fig-10. Precision-Recall Curve for Audio Stress Detection Model

B. Results of text-based stress detection.

The textual stress detection module is grounded on the RoBERTa-Large transformer architecture, being a contextual language understanding architecture. The model was adjusted on the Dreddit dataset, which has social media posts that were labeled stress and non-stress.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

RoBERTa model can comprehend contextual word to word relations and produce embeddings with linguistic information on stress and emotional states.

Table 5. Performance of Text Stress Detection Model

Accuracy	Precision	Recall	F1-Score
86.8%	0.87	0.89	0.88

These results indicate that transformer-based language models are effective in identifying psychological stress patterns in textual communication.

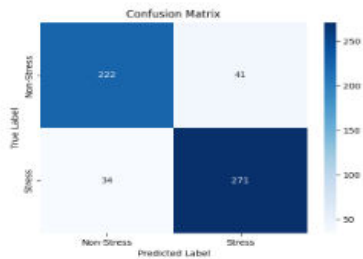


Fig-11. RoBERTa Text Stress Detection Model Confusion matrix.

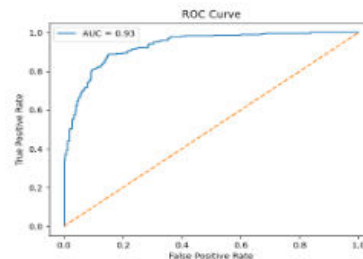


Fig-12. ROC Curve for RoBERTa Text Stress Detection Model

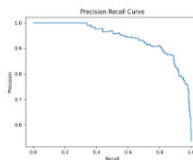


Fig-13. Precision–Recall Curve for Text Stress Detection

C. Cross-Dataset Evaluation

To further assess how well the text-based model can perform, the RoBERTa classifier was also tested on the speech transcripts produced by the Whisper speech recognition model on the RAVDESS dataset. The model is not as good when used with the RAVDESS dataset because the dataset is filled with scripted emotional words as opposed to actual textual materials that might be related to stress

Table 6. RoBERTa Performance on RAVDESS Transcripts

Model	Dataset	Accuracy
RoBERTa	RAVDESS Transcripts	63%

The lower accuracy is expected because the textual content of scripted emotional speech differs significantly from the real-life stress-related posts present in the Dreddit dataset.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

D. Hybrid Multimodal Stress Detection Results

The speech model and text model are combined with a multimodal fusion mechanism so as to enhance the performance of the stress detector.

The hybrid model combines:

- Acoustic characteristics derived out of speech signals.
- Contextual linguistic features obtained as text transcripts.

Such a mixture allows the system to receive the complementary information on both modalities.

Table 7. Experiments of Hybrid Multimodal Stress Detection Model.

Model	Dataset	Accuracy	Precision	Recall	F1-Score
Hybrid CNN-LSTM + RoBERTa	RAVDESS	0.9670	0.9610	0.9645	0.9721

The findings prove the effectiveness of the combination of acoustic and textual representations in enhancing the effectiveness and dependability of stress detection.

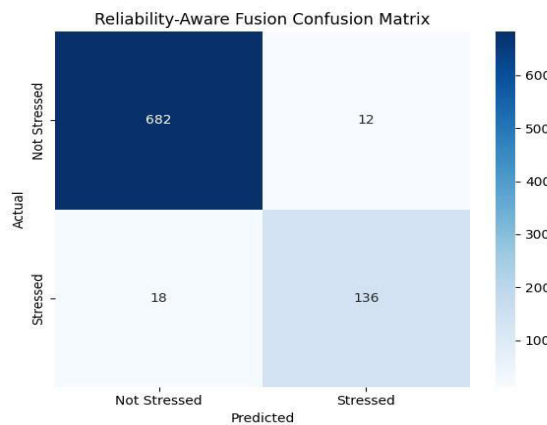


Fig-14. Confusion matrix of Hybrid (Audio + Text) model

E. Comparison with Existing Methods

Table 8. Comparison to Existing Stress Detection Methods.

Approach	Dataset	Model Type	Accuracy
SVM	Emotional Speech	SVM	78%
DNN	Speech Dataset	DNN	83%
CNN	RAVDESS	CNN	85%
Deep Learning	Speech Dataset	Deep Learning	90%
RoBERTa (Text Model)	Dreaddit	Transformer	86.8%
CNN-LSTM (Audio Model)	RAVDESS	Speech Model	96.11%
Proposed Hybrid Model (Audio + Text)	Multimodal	CNN-LSTM + RoBERTa	96.70%

The proposed CNN-LSTM + RoBERTa model has better accuracy than a number of traditional machine learning and deep-learning based models that are reported in the past studies



International Journal of Innovative Research in Computer and Communication Engineering (IJRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. DISCUSSION

The experimental findings prove the efficiency of the suggested multimodal model of detection of psychological stress. First, the CNN-LSTM speech model effectively deals with acoustic features of emotional stress in speech examples. Second, the text model created by RoBERTa is effective in the detection of linguistic indicators of psychological pressure in textual communication. Nevertheless, the text model exhibits comparatively poorer performance when used on speech transcripts produced off the RAVDESS dataset. This is not surprising given that the data used is scripted emotional sentences as opposed to actual stress-related expressions in the real-world.

The suggested multimodal system overcomes this weakness by combining speech and written material. By integrating acoustic and linguistic varieties, the system gains a more profound comprehending of the emotional stress patterns. In general, the experimental outcomes indicate that the suggested multimodal stress detection model offers a strong and efficient methodology of recognizing psychological stress in multimodal human communication.

VI. CONCLUSION

In this paper, a multimodal deep learning system has been introduced to identify psychological stress by integrating speech-based and text-based analysis techniques. The proposed research study will enhance the precision of automated stress detection systems with an increased amount of information in the form of speech and text-based communication media. The results of the experiment are that the speech-based CNN-LSTM model has already achieved an accuracy of 96.11 percent on the RAVDESS dataset. RoBERTa text model has been evaluated on Dreddit dataset and recorded an accuracy of 86.8 percent with a value of ROC-AUC 0.9315. Another experiment of the power of the textual model was to use the RoBERTa classifier on speech transcripts produced through the Whisper speech recognition model on the RAVDESS dataset. This model was also reported to be less accurate at 63, and this is not surprising because the data used in the RAVDESS dataset is scripted and associated with emotion and not with textual signs of stress that are prevalent in social media datasets.

The logic behind the proposed structure is that when the acoustic properties obtained using speech model are combined with the contextual and linguistic properties obtained using text model, then a more holistic picture of the stress signs in a person can be gained. Combining the two modalities, the resulting hybrid multimodal model reached an end-result accuracy of 96.70% and showed better results than unimodal models. Judging by the received experimental data, the combination of various modalities contributes greatly to the stability and effectiveness of automated psychological stress sensors.

REFERENCES

- [1] S. Cohen, D. Janicki-Deverts, and G. E. Miller, "Psychological stress and disease," *Journal of the American Medical Association*, vol. 298, no. 14, pp. 1685–1687, 2007.
- [2] J. Healey and R. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.
- [3] Z. Zeng, M. Pantic, G. I. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.
- [4] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
- [5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [6] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.
- [7] H. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.
- [8] S. Livingstone and F. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS)," *PLOS ONE*, vol. 13, no. 5, 2018.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- [9] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
- [10] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," *Proceedings of the International AAAI Conference on Web and Social Media*, 2013.
- [11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.
- [12] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*,
- [13] A. McEwen, "Protective and damaging effects of stress mediators," *New England Journal of Medicine*, vol. 338, pp. 171–179, 1998.
- [14] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH emotion challenge," *Proceedings of Interspeech*, pp. 312–315, 2009.
- [15] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," *Proceedings of ACM Multimedia*, 2010.
- [16] D. O'Shaughnessy, *Speech Communications: Human and Machine*. IEEE Press, 2000.
- [17] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," *Proceedings of Interspeech*, 2014.
- [18] A. Sarkar et al., "A review of speech emotion recognition using deep learning," *IEEE Access*, vol. 8, pp. 11171–11186, 2020.
- [19] R. Turcan and K. McKeown, "Dreaddit: A Reddit dataset for stress analysis in social media," *Proceedings of the EMNLP Workshop on Computational Linguistics and Clinical Psychology*, 2019.
- [20] A. Radford et al., "Robust speech recognition via large-scale weak supervision," *OpenAI Whisper*, 2022.
- [21] V. Bhandari, "Generating log-mel spectrogram using librosa," *Signal Processing StackExchange*, 2021. Available: <https://dsp.stackexchange.com/questions/75017/generating-log-mel-spectrogram-using-librosa>
- [22] R. Madhubala, K. R. Akhila, P. S. G. Aruna Sri, B. Madhav Rao, and A. Deepa, "Enhancing Cybersecurity with Intelligent AI and Machine Learning using AutoML Techniques," *International Journal of Applied Mathematics*, vol. 38, no. 4s, 2025.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details