International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Multimodal Human Stress Detection Using Deep Learning

Shweta Madale
Department of E&TC, Pimpri Chinchwad College of Engineering and Research, Ravet,Pune,India
shweta.madale_entc21@pccoer.in

Sanjana Deshpande
Department of E&TC, Pimpri Chinchwad College of Engineering and Research, Ravet, Pune, India
sanjana.despande_entc21@pccoer.in

Jayshri Birajdar
Department of E&TC, Pimpri Chinchwad College of Engineering and Research, Ravet, Pune, India
jayshri.birajdar_entc22@pccoer.in

Mrs. Dipali Dhake
Department of E&TC, Pimpri Chinchwad College of Engineering and Research, Ravet, Pune, India
dipali.dhake@pccoer.in

Abstract-

Stress is one of the most severe concerns in modern life, as it has become the integral part of each and everyone. Stress is the response given by the human body to the challenges and threats that can affect both the mind and body. Early detection of stress plays the crucial role in promoting the timely intervention. So, this project aims to detect stress with multiple data inputs using the deep learning models. Further the methodology of this continues with the pre-processing, feature extraction, data fusion, designing and training a deep learning architecture, neural network and evaluating its performance in terms of binary output, whether the person is stressed or not. Overall, this research contributes to the advancement of stress detection methodologies, offering a promising approach for early intervention and support in mental health care settings.

Keywords-CNN, LSTM, Transfer learning, MFCC , Decision level Data fusion.

Introduction

Data in multimodal carries the different information. Therefore, multimodal learning is described as the context of training the deep learning model using the multiple modalities of data, such as text, audio, or images. Unimodal models are trained using only using one type of data therefore they fail in terms of higher accuracy. We humans too make the decision after analysing the different sense input so does our machine should do the same. Stress detection using multiple modalities such as voice, facial expressions, Electroencephalogram (EEG), ECG helps increasing the outcome accuracy and to help the professionals to make accurate decision for interventions. In this project the use of two different modalities is done, such as EEG and speech for human stress detection. Selection of this modalities is done based on their importance in the stress detection process. EEG records electrical activity in the brain and can indicate various states of arousal, attention, and cognitive load, which are often linked to stress. Speech is another rich modality for stress detection. Under stress, the speech may show change in pitch, speaking speed etc.

The following segment presents an overview detailing the central themes and critical aspects discussed in this paper:

• The paper presents a stress detection system using a Deep learning model which is LSTM by utilizing the EEG dataset and the speech dataset.

• We delve into the significance of EEG brainwave and speech signal in detecting stress levels effectively.

• We review previous studies on various modalities for stress detection, highlighting methodologies, signal processing techniques, and the different types of deep learning methodologies.

• The results of our work demonstrate the accuracy and the F1 score of LSTM for stress detection for humans.

The rest of the paper is organized in the following manner:

Section II consists of information about Electroencephalogram signals, the various brain frequency bands and for Speech signal very basic features of audio to recognize are considered to be duration, MFCC, energy and pitch. Section III provides the literature review from the previous works that are related to the current work. Section IV is the methodology which we have implemented in our work. Section V provides the results and discussions about the overall work. Lastly, section VI presents the conclusion and future scope of the performed work.

By exploring novel methodologies, leveraging state-of-the-art technologies, and addressing existing challenges, we aim to develop a robust and reliable system capable

of accurately assessing stress levels from cues.

I. EEG Signals

EEG stands for electroencephalogram. Much electrical activity happens inside our human brain and this can be recorded and measured by the technique called EEG. It helps us detect electrical activities in the brain, including tumors, sleep disorders, brain injuries, and stress. EEG helps to study and indicate different brain states, such as wakefulness, sleep, and levels of alertness [8][9].

In Fig.1 EEG signal bands in different frequencies are given and it shows the different stages when they occur which is discussed further in this section. They significantly contribute in the detection of the stress level of a person. The waves are classified as Gamma, Beta, Alpha, Theta and Delta waves [14].

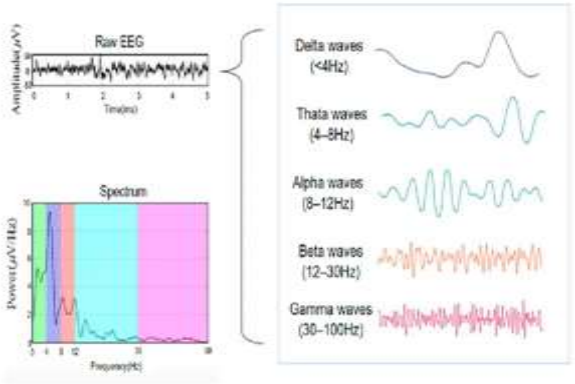

Fig. 1 EEG Brainwave

II. Speech Signals

Mel-Frequency Cepstral Coefficients (MFCC) depending on a linear cosine transformation (CT) of a joint angle Log power spectrum calculated on a non-linear Mel frequency scale, and it is also referred to as the short's spectrum Term regulation of the sound or audio. That is, any sound vocabulary created by humans is determined by the vocal tract shape, such as tongue, teeth, lips, etc. The envelope of the time power spectral density of the audio signal is representative of the vocal tract and MFCC, defined as the coefficients that comprise the Mel-frequency cephstrum and accurately represent this envelope.

III. Data Fusion

Data fusion can be described as the technique which involves combining the multiple sources for producing more consistent, accurate decision making capabilities of the model that could not be achieved by the single source alone.

Data Fusion can be done in three stages .

- **EARLY FUSION-** Fusing the data just after the feature extraction. This done converting the different modalities data into the single vector and simply concatenating the data.
- **LATE FUSION** – In the late fusion, each modal is train differently before passing it to the classifier. This help considering very single feature from each of the modals.
- **INTERMEDIATE FUSION** – In this method data fusion happens at different levels of Data Fusion , making it the more flexible.

IV. Literature Survey

With the emerging technologies, there are various computational methods that have been proposed over the years with the aim to characterize and detect human stress. From the sensors-based stress detection into the hospital to online stress detection at home using various application is possible now. In the following literature survey, a comprehensive examination of existing research on detection of human stress by using the different modalities i.e. Electroencephalogram signals, facial, and voice are presented. This survey observes diverse methodologies, signal processing techniques, and various deep learning approaches employed in previous studies to decode neural responses indicative of stress. The studies are done from the research papers that uses the single modal as well as multiple modals.

The work summited by Seo et al. (2023) [1] investigated using a deep learning method to identify stress related to work using signals from different sources. Study explores the integration of various signals to enhance the accuracy of stress detection, providing valuable insights for workplace stress assessment.

With the accuracy of 73%. Researchers Malviya and Mal (2022) [2] suggest a new method for identifying stress using EEG signals, combining deep learning techniques in an innovative way. The research contributes to the field by introducing an innovative model that combines different deep learning architectures such as CNN, GRU to improve the effectiveness of stress classification from EEG data, providing the accuracy up to 88%.

CNN-based facial recognition using the Kaggel dataset is done [3]. As the dataset is imbalance, so they employed the oversampling and under sampling along

with the normalization. Ater the 32 epochs are completed the model providing the 83% accuracy on the training data and 73.5% accuracy on the testing data.

Researchers Wan-Ting Chew, *Siew-Chin Chong, Thian-Song Ong and Lee-Ying Chong[4] experimented the various CNN architectures such as ResNet 50 , MobileNet V2 , Sequential model for the stress detection from the facial expression. Using the FER2013 dataset they obtained the accuracy of 50% for ResNet, 53% for the sequential model and the highest accuracy with the ESCNN.

The research done by Ankit, Mohd Akbar, Brijesh [6] firstly, frequency coefficients are extracted from speech using FFT algorithm. Then after the preprocessing the LSTM is used for the feature extraction. Lastly for the binary classification VGG 16 is implemented, providing the accuracy of the 98%.

The research contributed [7] [8] extracted the MFCC (Mel frequency cepstral coefficients) from the RAVDESS dataset extracted from the Kaggel. And then the CNN architecture is used to for the final binary classification. The CNN manages to provide the 76.8% of the accuracy.

In this paper, Cristian Paul Bara, Michalis Papakostas, and Rada Mihalcea contributed and presented how each module can be trained and reused independently and how different combinations of the modules can lead to different ways of combining modalities, each coming with its own advantages and disadvantages. They used the customise multimodal dataset MuSe for the research purpose.

The authors [7] proposed a hierarchical approach to feature extraction and fusion using the CNN architecture. This involves extracting features from ECG and EDA signals at different levels of granularity, from low level raw signals to high-level abstract representations. By combining these features at various levels, the model can learn more robust and discriminative representations.

The[10] 2023 study by RADHIKA KUTTALA 1 , RAMANATHAN SUBRAMANIAN 2 and VENKATA RAMANA 3 introduced multimodal Stress Detection Using hierarchical CNN-based feature fusion. This approach process the physiological (like EEG) and behavioral signals (like voice and facial expression)for stress detection. It has integrated feature at various levels-early, intermediate, and late stages- to capture the stress signatures. It had outperformed the unimodal and non-hierarchical CNNs.

V. Methodology

The framework of the methodology for the stress detection can be seen in the flow diagram shown below in Fig. 2. This chart outlines a process for detecting stress using a combination of brain, speech, and facial data. It starts with three inputs: EEG (brainwave) signals, speech, and facial expressions. Each of these inputs goes through a preprocessing step:

1. EEG: The brain signals are cleaned up through techniques like bandpass filtering and removing artifacts (unwanted noise).

2. Speech: The raw speech signals into meaningful features that can help the model understand and predict patterns.

After preprocessing and feature extraction, individual classifier is used to give the single probability output from each model, all the data (EEG, speech) is fused together. In the data fusion the average of the probabilities is calculated for the decision. Next, the model performs a classification to determine whether stress is present or not. If stress is detected, the system provides feedback. If no stress is detected, the model adjusts its weights in a training loop to improve accuracy over time. This loop of feedback and adjustment helps refine the system's ability to detect stress more accurately in future instances.

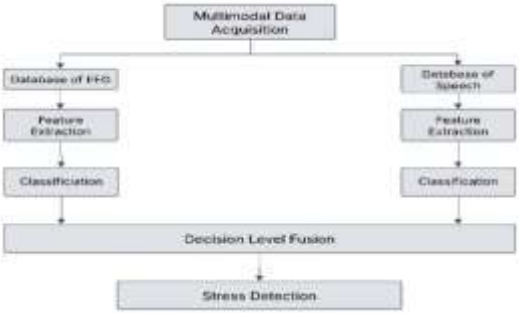

Fig 2. Flow chart of proposed system

1. Data Acquisition and Preprocessing

1.1. Data Collection:

Datasets for EEG and Speech signal is obtained. The dataset of EEG contains the amplitude values of the signals, it has around 200 to 300 amplitude values for further processing.

For Speech dataset it contains the .wav audio files. The data from the dataset is labelled into stressed and non-stressed. As the labelled dataset make the training of data quiet easy.

1.2. Pre - processing:

In EEG dataset to reduce the noise from the signal pre-processing of the data in done before the feature extraction so as to increase the accuracy of the output. The implemented system includes standardization, windowing, labelling, dataset Split.

EEG signal had noise signal and the varying amplitude because of the internal and external error, standardization is performed to transform the EEG signal to have **zero mean** and **unit variance** (μ=0, σ=1). In windowing, the EEG signal is divided into smaller, fixed-length segments (**windows**). Data is split into 80:20 for training and testing.

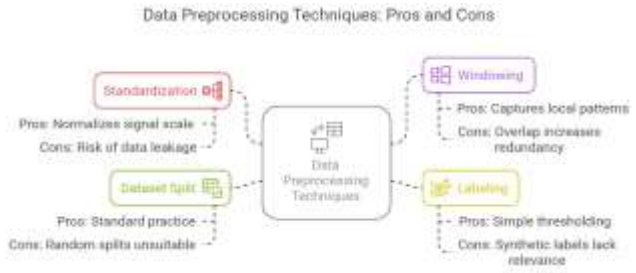

Fig 3. Data Preprocessing Techniques

Pre-processing is done for capturing physiologically relevant stress markers in human speech. Sampling of the data at the rate of 22.05khz- a sampling rate that optimally preserves relevant frequencies (85 Hz - 11.025 kHz). The waveform is then normalized in range [-1,1] maintaining consistent signal scaling across recordings. Files containing "stress" identifiers receive class label 1 (stressed) while others are labelled 0 (non-stressed), establishing a binary classification framework

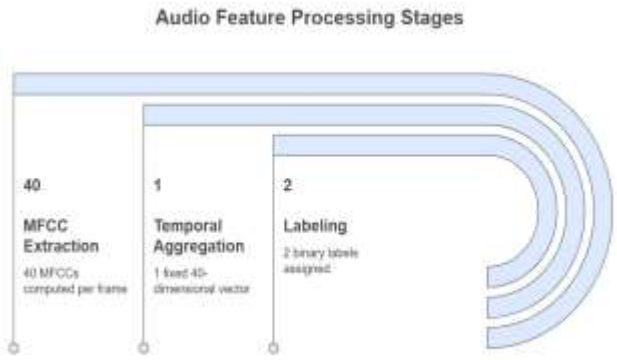

FIG 4. Speech Feature Processing

2. Feature Extraction:

In EEG signal from each window the energy is calculated.

Energy = Sum of squares of all data points.

The threshold for "high energy" is the **median energy** of the entire EEG signal. Then the data is labelled into high energy and low energy. If the window energy is greater than the median, it is labelled as 1 and if the energy is less than the median, it is labelled as 0.

MFCC pipeline is used for the feature extraction from the speech by creating the sample windows. To handle the variation in the recording, temporal aggregation computes the mean value of each coefficient across all the frame this preserve the spectral characteristics.

3. Classifier:

The model architecture of the EEG consists of the LSTM with the 64 unit, the dropout of the 30% to avoid the overfitting. Dense layer of 36 neuron and ReLU activation function to add non-linearity for feature transformation and then the dense layer of 36 neuron and softmax function, output layer for binary classification.

Training of the data for 10 epochs with 32 samples per batch and validating on 20 percent test data.

4. Data Fusion:

The probability output from the both EEG and the Speech model is fused in the data fusion part.

The average layer combines the probability output of each model for final predictions.

The[10] 2023 study by RADHIKA KUTTALA 1 , RAMANATHAN SUBRAMANIAN 2 and VENKATA RAMANA 3 introduced multimodal Stress Detection Using hierarchical CNN-based feature fusion. This approach process the physiological (like EEG) and behavioral signals (like voice and facial expression)for stress detection. It has integrated feature at various levels-early, intermediate, and late stages- to capture the stress signatures. It had outperformed the unimodal and non-hierarchical CNNs.

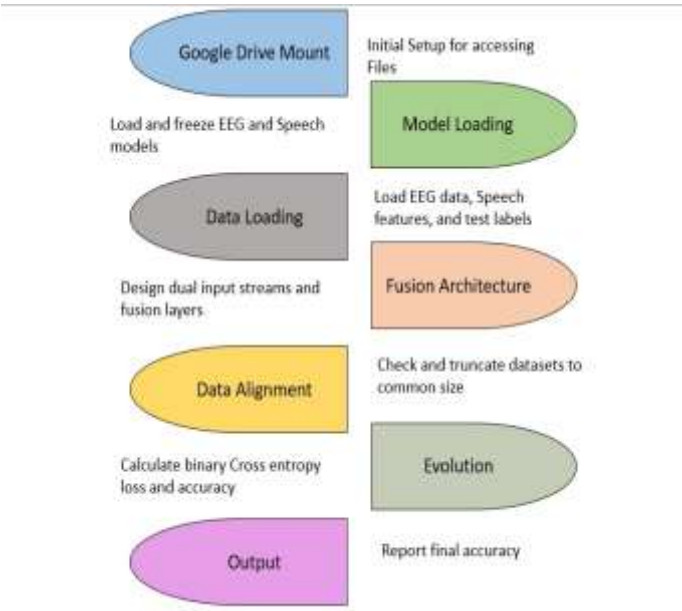

Fig 4. EEG and Speech Data Fusion Process

VI. Result and Conclusion:

The result output the binary classification, whether the person is stressed or not- stressed with the accuracy of 87.25% on the validation data from both EEG and Speech model.

This study explores the various deep learning architecture like LSTM for implementing the project. Using the different pre-processing techniques on each the modal and then features extraction before passing to the classifier is effectively studied and carried out. By using the different datasets the models are built to carry out the task of efficient stress detection in to the binary form. However, selecting the right classifier remains a challenge due the data collection, data synchronization and data fusion like tasks. Though implementing Multimodal Human Stress detection with the highest accuracy will be the ultimate aim or future scope, with any of the classifier.

VII. References:

1. W. Seo, N. Kim, C. Park, and S.-M. Park, “Deep learning approach for detecting work-related stress using multimodal signals,” IEEE Sensors Journal, vol. 22, no. 12

2. L. Malviya and S. Mal, “A novel technique for stress detection from EEG signal using hybrid deep learning model,” Neural Computing and Applications, vol. 34, no. 22, p. 19819–19830, 2022.

3. Enhancing Emotion Detection Through CNN-Based Facial Expression Recognition 5 Jinyang Wang Johnbapst Highschool, Bangor, ME 04401, USA (2020)

4. Facial Expression Recognition Via Enhanced Stress Convolution Neural Network for Stress Detection Wan-Ting Chew, *Siew-Chin Chong, Thian-Song Ong and Lee-Ying Chong A IAENG International Journal of Computer Science.

5. Video-Based Stress Detection through Deep Learning Huijun Zhang * , Ling Feng, Ningyun Li, Zhanyu Jin and Lei Cao (2020)

6. Identification of psychological stress from speech signal using deep learning algorithm Ankit Kumar a , Mohd Akbar Shaun b , Brijesh Kumar Chaurasia (2024)

7. Stress and Anxiety Detection through Speech Recognition Using Deep Neural Network Divyashree P1 , Ghanavi Yadav A2 , Namratha Jayadev3 , Prajwal B4 , Sharmila Chidaravalli(2022)

8. Stress Detection through Speech Analysis using Machine Learning Dr. S. Vaikole, S. Mulajkar, A. More, P. Jayaswal, S. Dhas (2020)

9. A Deep Learning Approach Towards Multimodal Stress Detection Cristian Paul Bara, Michalis Papakostas, and Rada Mihalcea(2020)

10. Multimodal Hierarchical CNN Feature Fusion for Stress Detection RADHIKA KUTTALA 1 , RAMANATHAN SUBRAMANIAN 2 , (Senior Member, IEEE), AND VENKATA RAMANA MURTHY ORUGANTI 1 , (Senior Member, IEEE) (2023)