Facial and Speech Emotion Detection for Stress Monitoring

Mrs. S. Jansi Rani[1], Arulmoneesh. E[2], Dharaneesh. R[3] and Mahesh Veeraiah. R[4].
[1] *Assistant Professor, Sri Ramakrishna Engineering College, Coimbatore, India*
[2,3,4] *UG Scholar, Sri Ramakrishna Engineering College, Coimbatore, India*

Abstract **- Stress is now a common problem in modern life, frequently going unnoticed until it has an impact on one's physical and mental health. This study offers a multimodal stress detection framework that uses facial emotion analysis and speech recognition to more accurately predict stress levels. A refined DistilBERT model is used to analyse speech input into text before classifying emotions and mapping them to appropriate stress levels. In parallel, a Convolutional Neural Network (CNN) trained on the FER dataset is used to identify emotional states and MTCNN is used for face alignment in order to detect facial expressions. In both modalities, the suggested method shows good classification performance and has the potential to be included into real-time stress monitoring applications.**

Index Terms **- BERT, CNN, Emotion Classification, Facial Expression Recognition, FER Dataset, MTCNN, Multimodal Analysis, Speech Emotion Recognition, Stress Detection, Transformer Models.**

I. INTRODUCTION

Stress, a psychological reaction to challenging circumstances, is becoming more widely acknowledged as a significant element affecting mental and physical health. Given how quickly life is changing in the modern world, it is essential to identify stress levels early on in order to prevent long-term health problems like anxiety, depression, and heart disease. Self-reporting or physiological monitoring are frequently used in traditional stress evaluation methods, which might be invasive, time-consuming, or have a narrow scope.

Recent developments in deep learning and artificial intelligence have made it possible to create automated systems for recognizing emotions, which act as a proxy for stress. Anger, fear, and melancholy are examples of emotional states that are generally connected to increased stress levels, whereas joy and love are linked to lower stress levels. In order to improve the accuracy of stress level detection, this study presents a dual-modality approach that makes use of both speech and facial expressions.

In this work, a refined DistilBERT [3] model-a lighter and faster version of BERT [3]-that can comprehend contextual semantics in transcribed speech is used to perform speech-based emotion recognition. A Convolutional Neural Network (CNN) trained on the FER dataset is used to execute facial emotion recognition concurrently, after MTCNN [6] has been used for face identification and alignment. Emotional states are independently classified by both modalities and linked to Low, Neutral, and High stress levels. Building an effective, scalable, and non-intrusive stress detection algorithm that can be incorporated into wearable or mobile health applications is the aim of this study.

II. LITERARY SURVEY

This survey reviews recent research on stress detection using multimodal approaches, integrating facial expressions, speech, and physiological data. The studies highlight the use of deep learning models such as CNN, BERT, and MTCNN to improve accuracy and reliability in identifying stress indicators.

Rahee Walambe, Pranav Nayak, Ashmit Bhardwaj, Ketan Kotecha [1] This study presents a multimodal AI-based framework for stress detection, integrating data from facial expressions, posture, heart rate, and computer interaction. The system utilizes a fusion of heterogeneous raw sensor data streams, enabling it to detect stress with an accuracy of 96.09%. The authors demonstrate that combining multiple data modalities improves the reliability of stress detection, especially in real-time monitoring of IT professionals. The framework's strength lies in its ability to process and analyze various biometric signals and behaviors to predict stress, offering potential applications in workplace health and well-being.

Majid Hosseini, Morteza Bodaghi, Ravi Teja Bhupatiraju, Anthony Maida, Raju Gottumukkala [2]

The authors propose a multimodal learning approach for stress detection that integrates facial landmarks and biometric signals such as heart rate and skin conductivity. They explore both early-fusion and late-fusion techniques, where the former combines features from different modalities before feeding them into a model, and the latter does so after processing each modality independently. The study found that late-fusion achieved an accuracy of 94.39%, while early-fusion surpassed it with a 98.38% accuracy rate, highlighting the potential for multimodal integration in improving stress detection. The framework provides a promising solution to detect stress in real-time and in naturalistic settings.
Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, Dinesh Manocha [3] The M3ER framework combines facial, textual, and speech cues for emotion recognition, which are particularly important for detecting stress in a multimodal context. By employing a multiplicative fusion method, the system is able to dynamically emphasize reliable cues (e.g., speech tone or facial expression) and suppress less useful ones, achieving a mean accuracy of 82.7% on the IEMOCAP dataset. This framework emphasizes the importance of integrating different modalities, allowing for a more accurate understanding of human emotional states, including stress. The results highlight the efficacy of multimodal fusion in emotion recognition tasks, particularly in detecting subtle signs of stress.

Sarala Padi, Seyed Omid Sadjadi, Dinesh Manocha, Ram D. Sriram [4] This research introduces an emotion recognition framework based on neural networks that incorporates transfer learning techniques. The system adapts a ResNet-based model, which was initially trained on large-scale speaker recognition tasks, and fine-tunes it with a BERT model for processing textual data. The authors demonstrate that transfer learning, combined with BERT's language understanding capabilities, enhances the system's ability to classify emotional states, including stress, achieving state-of-the-art results on the IEMOCAP dataset. The research highlights the advantages of using pre-trained models for multimodal emotion recognition, significantly improving performance by leveraging both speech and textual cues.

Salman Mohammed Jiddah, Burcu Kuter, Kamil Yurtkan [5] This study proposes a method for detecting stress through compound facial expressions, analyzed using deep neural networks. The system processes combinations of basic facial expressions (such as anger, surprise, and sadness) to detect complex stress indicators. By using advanced deep learning techniques, the authors achieve encouraging results in detecting stress levels, demonstrating the potential of analyzing facial expressions as a reliable indicator of emotional states. The system also contributes to the field of emotion analysis by using compound facial features, offering a new approach to recognizing and classifying stress and other emotions.

III. METHODOLOGY

Speech emotion and face emotion are two different modalities that are intended to be used by the suggested system for classifying stress levels. A common stress scale with some pre-defined classes mapped to each modality after it has been processed independently. The model designs, training protocols, performance evaluation techniques, and data preparation are all covered in this section.

3.1 Speech Emotion Analysis Using DistilBERT

Speech-to-Text Conversion

A voice-to-text engine is used to convert the user spoken signals into text. This conversion makes it possible to classify emotions using NLP-based models.

Dataset Preparation

Transcribed sentences with emotional state labels make up the input data. A rule-based mapping is used to automatically correlate stress levels with emotions.

Model Training

Hugging Face's Transformers and Trainer API are used to refine DistilBERT [3], a distilled, lighter version of BERT [3], for sequence categorization. The tokenized dataset is used to train the model with output labels. To guarantee a balanced distribution across all classes, a stratified train-validation-test split is employed.

Evaluation Metrics

Metrics like accuracy, F1-score, and confusion matrix are used to assess the model after training. These measurements aid in evaluating performance across classes, particularly when it comes to differentiating between stress indications that are high and low.

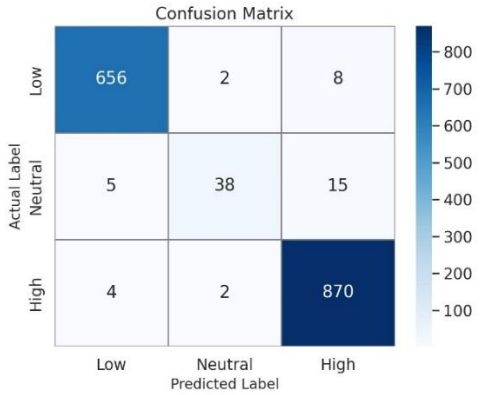

Fig 1 Confusion Matrix (DistilBERT)

3.2 Facial Emotion Recognition Using MTCNN + CNN

Face Detection and Preprocessing

Faces in video frames are detected using MTCNN [6]. Prior to categorization, it ensures consistency by performing face localization and alignment. Face photos that have been cropped and aligned are downsized to a common size (such as 48x48) that can be used as CNN [6] input.

Model Architecture

The FER2013 dataset is used to train a Convolutional Neural Network to categorize photos into fundamental emotion groups. For classification, the CNN [6] has multiple convolutional and pooling layers, which are followed by fully connected layers.

Emotion-to-Stress Mapping

The same stress markers used in the speech model are applied to each face emotion following emotion classification.

Model Evaluation

The efficacy of the CNN [6] model is evaluated using its training history and final accuracy on test and validation datasets.

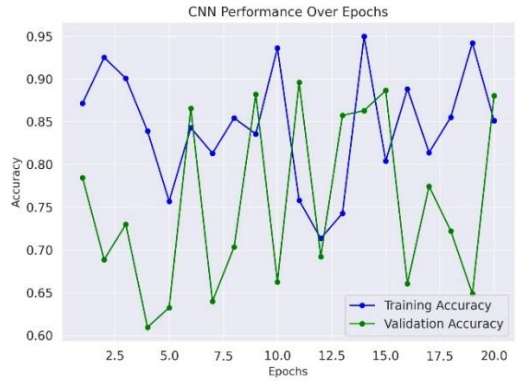

Fig 2 CNN Performance Graph

3.3 Unified Stress Mapping and System Flow

The project's ultimate objective is to develop a multimodal fusion system where both inputs contribute to a final aggregated stress score, even if the speech and facial models operate independently in current iteration. For the sake of clarity and accuracy, we address the modalities independently in this work.

The following phases make up the system architecture:

- Acquisition of Input (Video/Speech)
- Preprocessing (voice transcription, image face identification)
- Emotion Recognition (face CNN, speech DistilBERT)
- Final Stress Label Output
- Emotion-to-Stress Mapping

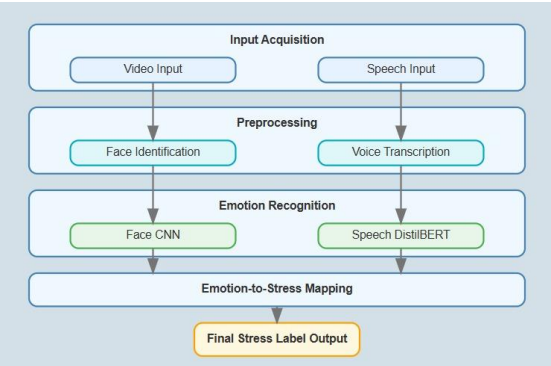

Fig 3 Flow Diagram

IV. RESULTS AND DISCUSSION

The experimental results are shown in this section along with an analysis of the effectiveness of the speech-based and facial emotion-based stress classification models.

3.1 DistilBERT Model Performance (Speech Emotion Analysis)

Based on transcribed speech, the refined DistilBERT [3] model performed well in identifying stress levels. On the test dataset, the model's classification accuracy was high for all given labels.

Among the important findings are:

- Accuracy: The model's total test accuracy was almost 95%.
- Confusion Matrix Analysis: Misclassifications were more frequent among certain groups, suggesting that emotional meanings occasionally overlap.

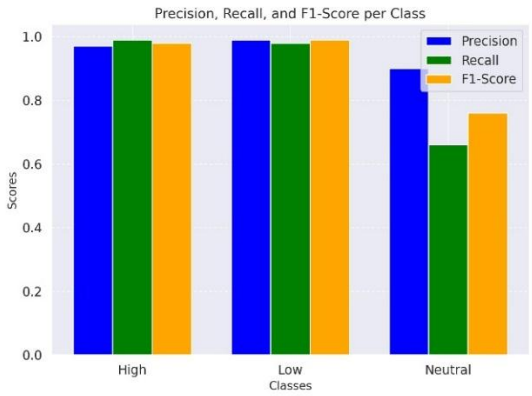

Fig 4 Bar Chart of Precision, Recall and F1-Score

3.2 CNN Model Performance (Facial Emotion Analysis)

The CNN [6] trained on the FER dataset also performed effectively in recognizing emotions and mapping them to stress levels.

Performance metrics include:

- Training Accuracy: 85
- Validation Accuracy: 87
- Loss Trends: The training and validation loss curves stabilized after a few epochs, indicating that the model successfully converged.

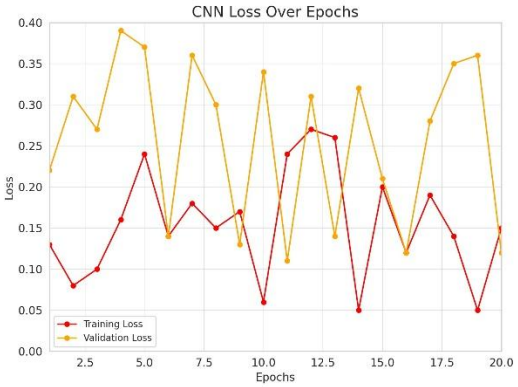

Fig 5 Training and Validation Accuracy Graph

Challenges observed:

Sometimes facial expressions like "Surprise" were misunderstood because they overlapped with other emotions in visual signals. Despite its widespread use, the FER dataset contains cropped and grayscale photos, which could hinder performance in real-world colourful or changing lighting circumstances.

3.3 Comparative Analysis and Limitations

Even though both models worked well on their own, robustness could be improved by a combined multimodal analysis (for example, by employing fusion techniques), particularly in cases where one modality fails or produces ambiguous results. Furthermore, the existing system lacks physiological sensor data and real-time feedback, both of which would increase the accuracy of stress assessment.

4. CONCLUSION

Even while both models performed well independently, a combined multimodal analysis (for instance, by using fusion techniques) could increase robustness, especially when one modality fails or yields unclear results. Additionally, the current system is devoid of real-time feedback and physiological sensor data, both of which would improve the precision of stress measurement.

REFERENCES

[1] A. S. M. Iftekhar, M. A. Rahman, and M. A. H. Akhand, “A deep learning approach for sentiment analysis of COVID-19 tweets,” IEEE Trans. Neural Netw. Learn. Syst., vol. 4, no. 5, pp. 1–6, Apr. 2023, doi: 10.1109/ECCE58298.2023.10100505.

[2] S. M. Jiddah, B. Kuter, and K. Yurtkan, “Stress detection through compound facial expressions using neural networks,” Eurasia Proc. Educ. Soc. Sci., vol. 35, pp. 1–9, 2024.

[3] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” arXiv preprint arXiv:1910.01108, Oct. 2019.

[4] Y. Jiang, B. Sharma, M. Madhavi, and H. Li, “Knowledge distillation from BERT transformer to speech transformer for intent classification,” presented at the Proc. Conf. on Spoken Language Understanding, National Univ. of Singapore, Singapore, 2021.

[5] B. Buyukoz, “Analyzing the generalizability of deep contextualized language representations for text classification,” arXiv preprint arXiv:2303.12984, Mar. 2023.

[6] R. Nachet and T. B. Stambouli, “Improved face recognition rate using convolutional neural networks,” Proc. 2022 2nd Int. Conf. New Technol. Inf. Commun. (NTIC), pp. 1-6, 2022.