Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 10s (2025)

A Multimodal Framework for Assessing Mental Health Strain among IT Industry Workers Using Machine Learning

Mr.Shailesh Kurzadkar [1], Dr.Vijay Bhandari [2], Dr.Anup Bhange[3]

[1]PhD Scholar, Department of CSE, Madhyanchal

Professional University, Bhopal,India

[2]Assosiate Professor, Department of CSE, Madhyanchal

Professional University, Bhopal,India

[3] Assistant Professor KDK College of Engineering, Nagpur, India

Article History:

Received: 19-03-2025

Revised: 24-07-2025

Accepted: 21-09-2025

Abstract:

The growing prevalence of work-related stress highlights the need for unobtrusive, continuous monitoring tools. This paper proposes a multimodal system for real-time detection of employees' emotional states during video conferences, with a focus on stress. The framework integrates Facial Emotion Recognition (FER), speech transcription using OpenAI's Whisper, NLTK VADER. The results from experimental show encouraging performance: Whisper gives accurate transcription, NLTK VADER sentiment analysis achieves great classification accuracy across five emotion categories, and multimodal fusion extents up to 88% accuracy in stress detection. This work founds a basis for real-time, automated valuation of employee well-being, allowing adaptive interferences and associate healthier workplace surroundings.

Keywords: Employee Feeling Detection, Stress Detection, Multimodal Fusion, Deep Learning, Natural Language Processing, Facial Emotion Recognition, Whisper Model, , Real-time Monitoring

1. Introduction

In current years, research on stress analysis has grown rapidly. Great levels of emotional stress have been connected to dangerous behavioral deviations in both men and women. Stress is basically the body's response to internal or external pressures, and it can negatively mark memory, decision-making, and daily functioning. Extended contact to acute stress may lead to worsening of both physical and mental health, possibly causing immune dysfunction, cardiovascular issues, depression, or other illnesses.[4] In modern society, stress has become widely main and severe. It can be measured through questionnaires or professional psychological consultations. However, since psychological assessments are subjective and momentary, they often produce inaccurate or misleading results and fail to meet the requirements of real-time stress detection[19].

Prolonged exposure to stressors is typically linked to negative physical and psychological results, including prolonged illness, burnout, and depression. This underlines the urgent need for continuous, reliable, and inconspicuous systems that can detect and monitor stress and overall emotional well-being at an early stage. Conventional valuation methods, such as self-report surveys, often suffer from subjectivity, troublesomeness, and bias. Advances in artificial intelligence and machine learning joint with the extensive use of digital communication platforms like video conferencing and real-time monitoring of mental states.

This study introduces a multimodal framework for detecting employees' emotional states, with a particular emphasis on stress, using data from video conferencing and from Microsoft's FER+ (Facial Expression Recognition Plus) Dataset. The system incorporates machine learning and natural language processing methods to analyze multiple data streams, including facial expressions, speech, and vocal tone. Its main objective is to deliver a complete view of employees emotional well-being, enabling early identification of stress and easing timely intervention. The paper details the system's architecture, preprocessing and analysis techniques across modalities, experimental design, and expected outcomes, supported by both implementation insights from a prototype notebook and findings from related research.

Changes in voice, facial expressions, and involuntary bodily functions often accompany stress. FER technologies powered by computer vision and artificial intelligence have grown rapidly in recent years. Facial expressions, produced by muscle movements play a important role in human communication. They spread a wide range of information, from subtle cues like an eyebrow raise in conversation to signals tied to basic survival. Factors such as lighting, pose, background, camera angle, sealing, and misalignment significantly influence FER performance. Effective FER depends on both perceptual system data and computations within the visual-perceptual system, supported by perceptual processes.

So far, maximum research has attentive on workflow design and feature extraction. FER has been explored from both academic and industrial scenarios, as it can disclose insights into personality, temperament, cognitive ability, and psychological disorders.

Facial expressions often trigger immediate responses and commonly mirror emotions. When facial muscles contract in response to a particular event or question, they can rapidly communicate significant information in both social and managerial interactions. Therefore, automatic FER methods are essential for computational systems to accurately evaluate an individual's emotional state.[33]

This study focuses on evaluating the transcription performance of Whisper, an Automatic Speech Recognition (ASR) model developed by OpenAI [37]. The tool is being used within a research project designed to create an automatic support system for diagnosing depression, where it serves to transcribe audio recordings from both depressed and healthy participants.

Given the increasing integration of technology into both personal and professional life, it is natural to reflect on its impact on communication and collaboration, as well as the potential

benefits of Interactive Artificial Intelligence in promoting and supporting mental health. The objective is to detect speech features that could indicate the presence of depression.

More precisely, our focus was on examining verbal behavior, which makes it possible to examine both the content and the linguistic style—two dimensions that may reveal cognitive patterns and emotional states (such as anger, anxiety, or sadness) that are prevalent or maladaptive in depressive disorders.

The collected audio recordings first had to be transcribed with Whisper, an Automatic Speech Recognition (ASR) model developed by OpenAI.

Generally, ASR systems gets acoustic input from a speaker via a microphone, process it using patterns, models, or algorithms, and generate an output—typically in text form. The performance accuracy of a speech recognition system is influenced by several factors, including whether it is speaker-dependent or independent, the use of discrete or continuous word recognition, vocabulary size, and environmental conditions. Whisper represents the newest advancement in a line of Convolutional Transformer End-to-End ASR models.

This study concentrates on Whisper's transcription of the "tale task." The rationale for this choice is that evaluating Whisper's performance required starting with a ground truth—the original fable text—to compare against its final transcriptions. Data from the Diary task have already been examined and discussed in other studies on Verbal Behavior Analysis, which have been submitted for publication and are currently under review. [37]

2. Related Work

To detect acute stress, Zhang et al. [2] introduced a deep learning framework that integrates voice, facial expressions, and ECG data in real time. Building on this, we developed a Temporal Attention Module (TAM) to pinpoint keyframes relevant to stress detection from facial expressions. The proposed approach requires only minimal preprocessing and eliminates the need for complex feature engineering. Our contributions can be summarized as a deep learning system for severe stress detection that combines voice, facial, and ECG modalities, incorporates TAM, and employs a matrix eigenvector-based fusion strategy, achieving 85.1% detection accuracy. By assigning different learnable weights to individual frames, the TAM captures distinctive temporal patterns in facial expressions associated with stress.

Baheti Reshma Radheshamjee and Kinariwala Supriya [13] proposed a system that identifies stress through in-person interviews, discussions, or similar activities, typically involving analysis of two or more individuals by an external evaluator. Another framework was introduced that leverages users' weekly social media activity, focusing on social interactions and tweet content to assess psychological stress levels. In this approach, each dictionary term is assigned a score ranging from –5 to +5. Both Naïve Bayes (NB) and Support Vector Machine (SVM) algorithms were applied for classification and prediction. To enhance accuracy, Word Sense Disambiguation (WSD) was employed using n-gram and skip-gram models. When combined with SVM, the use of n-gram and WSD achieved a precision of 65%.

Stress among office workers is increasing due to heavy workloads and job-related pressures. To mitigate negative effects on well-being, it is crucial to monitor and manage employees' stress levels regularly, enabling early detection. A review of the literature on workplace stress identification through multimodal measurements [24] highlighted key features and parameters across physical, social, and related data.

In [3], the study utilizes both primary and secondary datasets to investigate the impact of social and emotional factors on social media data and usage. Adopting a cross-sectional approach and combining qualitative with quantitative methods, the research focuses on building models for sentiment and emotion detection that can support stress management. These models are further evaluated using original data. To analyze user sentiments and emotions across various themes, Latent Dirichlet Allocation (LDA) is applied. Additionally, sentiment analysis is conducted through hybrid machine learning and deep learning models trained on a diverse collection of tweets.

A recently developed technique estimates the effective integration window of an arbitrary response by presenting segments of natural stimuli—such as speech—in two distinct random sequences, ensuring each segment appears within different contexts, surrounded by varying stimuli [36]. In another study, [37] proposed a system for categorizing song tones into appropriate classes using a supervised learning approach to analyze the tones of specific lyrics. The methodology is divided into four stages: data collection, data processing, tone class extraction with IBM Watson Tone Analyzer, and classifier-based tone classification.

Marian Pompiliu Cristescu et.al [38] developed a Python-based approach utilizing the BERT and VADER libraries to process and evaluate textual data. Their study identifies notable differences in sentiment polarity outcomes between these tools, with BERT exhibiting a tendency toward positivity and VADER producing a more balanced sentiment distribution. These results underscore the need to carefully choose sentiment analysis tools based on the characteristics of the text under examination. Ultimately, their research demonstrates both the effectiveness and significance of open-source tools in sentiment analysis, especially for improving continuous feedback systems in organizational settings.

3.Methodology

The proposed system for detecting employee emotional states employs a multimodal deep learning framework that analyzes video and audio data from video conferences. Its methodology includes sequential stages of data acquisition, preprocessing, feature extraction, and classification, applied to facial expressions, speech transcription, and text-based sentiment.

3.1 System Architecture Overview

The system's core design involves three primary analytical streams:

1. Facial Emotion Recognition: Directly from video frames and from FER + Datasets
2. Speech Transcription Whisper Model: From the audio track of the video.
3. Text Sentiment Analysis: Applied to the transcribed speech.

3.2 Data Acquisition and Preprocessing

3.2.1 Facial Expressions Preprocessing

Video frames are extracted and converted to RGB (or BGR, depending on the library) to ensure compatibility with emotion detection models. For face detection and alignment, the Facial Emotion Recognition (FER) library is initialized with mtcnn=True, enabling the use of Multi-task Cascaded Convolutional Networks (MTCNN). This approach isolates facial regions by removing background information, thereby reducing noise and visual clutter.

Figure 1 illustrates the generic pipeline of FER's stepwise working flow, which is described in this part.

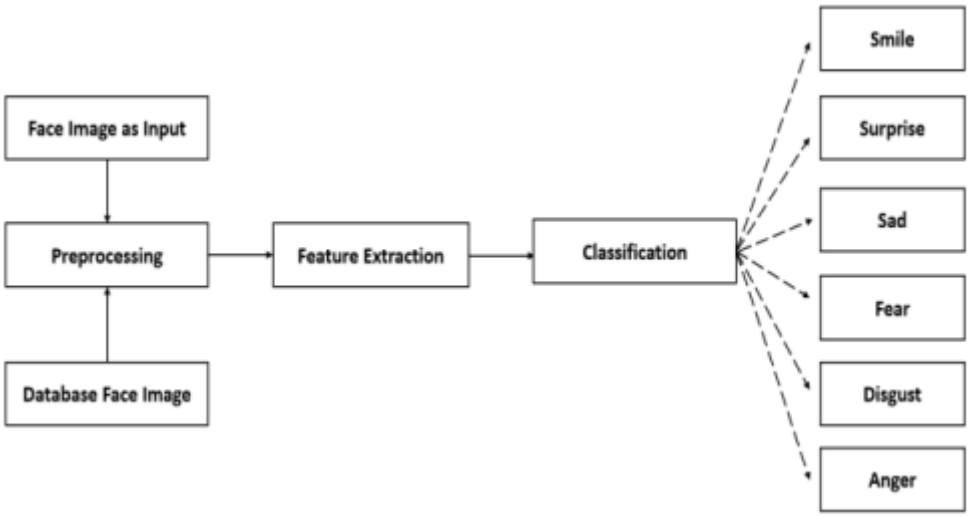

Figure-1

Preprocessing and data collecting via visual sensors are crucial phases. Usually, the information is obtained from a variety of sources, including surveillance, mobile phones, and Pi Cam devices.,cameras.

Facial detection is another name for region of interest (ROI) detection, which in this case refers to the face. AI-based methods are used for ROI detection, which finds and identifies faces in pictures.

These techniques have been widely used in a number of applications that entail tracking or surveillance, including security, law enforcement, entertainment and personal safety. A system called facial emotion recognition analyzes emotions from a variety of sources, including images and videos.

Human emotions can be inferred from facial expressions, which are nonverbal communication methods. Researchers in the fields of psychology (Ekman and Friesen 2003; Lang et al. 1993) and human-computer interaction (Cowie et al. 2001; Abdat et al. 2011) have been interested in decoding such emotion displays for decades. The development of FER

technology has recently been significantly influenced by the widespread use of cameras as well as advancements in machine learning, biometrics analysis, and pattern recognition.

The photos or videos that are used as input to FER algorithms come from a variety of sources, including surveillance cameras, cameras positioned near storefront advertising screens, social media, streaming services, and individual devices.

The overall process of face emotion recognition consists of three phases: Pre-processing, face detection, and sentiment classification. In the pre-processing phase, the dataset is prepared to work with generalized algorithms and generate efficient results. The face detection phase involves detecting faces in real-time captured images

3.2.2 Speech Audio Preprocessing

Audio is separated from the video stream for transcription. For broader audio analysis, such as tone evaluation, preprocessing involves applying pre-emphasis to the signal, followed by framing with a 30 ms Hamming window and a 15 ms overlap.

Whisper for language identification

The Whisper model follows a straightforward end-to-end design based on an encoder-decoder Transformer. Audio input is divided into 30-second segments, transformed into a log-Mel spectrogram, and processed by the encoder. The decoder is then trained to generate the matching text output, incorporating special tokens that enable the model to handle tasks like language detection, phrase-level timestamping, multilingual transcription, and translation of speech into English.[38]

Accent Identification and Language masking

Accents represent differences in pronunciation, intonation, rhythm, and speech patterns that emerge from geographical, social, cultural, or linguistic influences. Recognizing and interpreting accents is useful for a range of applications, including speech recognition, speaker diarization, language learning, and sociolinguistic research.

In the framework the process of finding and segmenting multiple languages within a multilingual audio recording—accent detection plays an important role. Accents can act as an additional signal to distinguish between languages and to detect language transitions, even within code-switching segments. Combining accent-based features with other acoustic and phonetic cues can significantly improve the accuracy of language segmentation and identification, especially in linguistically diverse environments.[40] Throughout the paper, the terms attention mask, language masking, and language selection are used interchangeably.

Whisper model uses Transformer-based sequence-to-sequence learning algorithm.

In Whisper, input = audio spectrogram, output = text tokens. The Transformer is the neural network architecture used to perform this mapping.

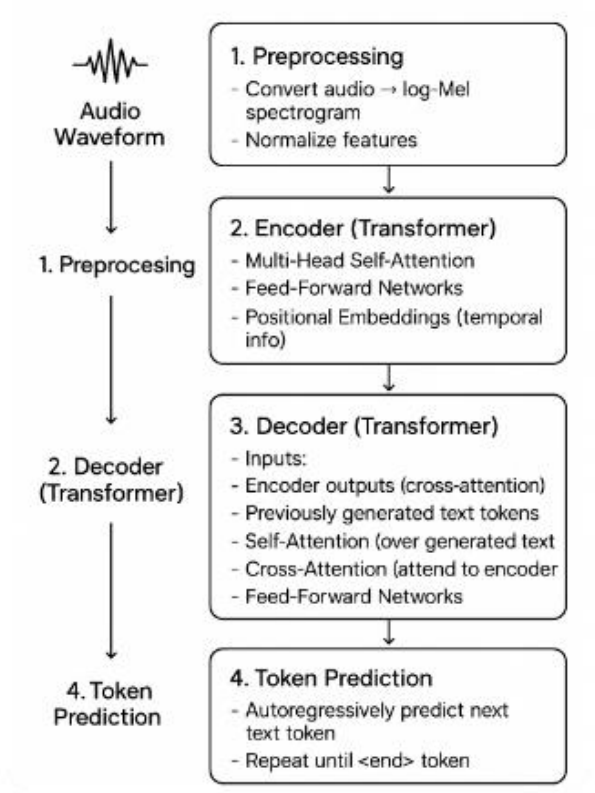

Figure-2: Whisper Seq2Seq Architecture Flow

Encoder

- Input sequence: log-Mel spectrogram (from audio).
- Send over multiple layers of **self-attention** + **feed-forward networks**.
- Captures relations through time and frequency (e.g., phonemes, rhythm).
- Outputs a sequence of contextualized embeddings.

Decoder

- Takes encoder outputs + previously generated text tokens.
- Uses **self-attention** (over generated text so far) and **cross-attention** (attend to encoder outputs).
- Autoregressively generates next tokens until the sentence ends.

Attention Mechanism

- Core of the Transformer.
- Helps the model learn **which parts of the input are relevant** for predicting the current output token.

3.2.3 Transcribed Text Preprocessing

Succeeding transcript, the text is processed through several NLP preprocessing steps:

- **Language Detection:** Achieved with langid.classify to identify the text's language.
- **Tokenization and Lemmatization:** Splitting text into tokens and converting words to their root forms.
- **Stop Word Removal:** Eliminating common terms (e.g., "is," "a," "the"), with added support for Hindi stopwords.
- **Keyword Extraction:** Applying the RAKE (Rapid Automatic Keyword Extraction) algorithm to identify key phrases and terms by filtering out stop words and using delimiters.

3.3 Feature Extraction

3.3.1 Facial Expressions

The 'FER' library detects faces and extracts associated emotions (e.g., 'sad', 'angry', 'neutral') along with their confidence scores from video frames

3.3.2 Speech Audio

OpenAI's Whisper model is employed for Automatic Speech Recognition (ASR),

transcribing the audio into a complete textual transcription

For general audio analysis, features extracted can include energy density and Mel spectro-gram

3.3.3 Transcribed Text

VADER Sentiment Analysis:

The nltk.sentiment.vader.SentimentIntensityAnalyzer produces polarity scores over four categories: neg (negative), neu (neutral), pos (positive), and a compound score. The compound value ranges from -1 (strongly negative) to +1 (strongly positive), with scores between -0.5 and 0.5 generally considered neutral.

VADER is a lexicon and rule-based sentiment analysis tool.

Lexicon-based Scoring

- VADER has **already defined dictionary (lexicon)** of words, each associated with a **valence score** (ranging from –4 to +4).
- Example: *"great" = +3.1, "bad" = –2.5, "happy" = +2.7.*
- When analyzing text, each token is matched with the lexicon and its sentiment score is retrieved.

In VADER's lexicon-based scoring, the algorithm uses a valence summation formula to combine word-level sentiment scores into a text-level score.

For each token *w* in the text:

valence(w)=score(w)+modifiers−negations+emphasis

All adjusted word scores are summed up

$$S = \sum_{i=1}^{n} valence(w_i)$$

where *n* = number of sentiment-bearing words.

To keep the sentiment score in a bounded range (–1 to +1), VADER applies a normalization formula:

$$compound = \frac{S}{\sqrt{S^2 + \alpha}}$$

All adjusted valence scores are summed and then normalized with

$$compound = \frac{S}{\sqrt{S^2 + 15}}$$

In VADER's rule-based heuristics, it applies adjustment formulas depending on context (negation, intensifiers, punctuation, etc.).

1. Degree Modifiers (Intensifiers / Dampeners)

Each modifier word (e.g., "extremely", "very", "slightly") has a scalar value in the VADER lexicon.

If modifier precedes a sentiment word:

$$valence' = valence + (valence \times modifier\_scalar)$$

2. Negation Handling

If a negation word (e.g., "not", "never", "isn't") occurs within 3 tokens before a sentiment word:

$$valence' = valence \times -0.74$$

3. Punctuation Amplification

Each exclamation mark (!) increases intensity by +0.292, capped at 4 marks.

$$valence' = valence + (n\_exclamations \times 0.292)$$

4.Capitalization Emphasis

$$valence' = valence + (valence \times 0.733)$$

3.4 Classification

3.4.1 Facial Expressions

The ‘FER‘ library integrally does emotion classification on detected faces, providing counts of various emotions

3.4.2 Transcribed Text Sentiment

NLTK VADER Sentiment Intensity Analyzer outputs polarity scores for general sentiment.

3.4.3 Multimodal Fusion for Overall Feeling/Stress

A real-time deep learning framework is introduced for analyzing facial expressions. The fusion approach employs probability data from each modality’s stress level is combined into a stress information matrix.

3.5 FER Algorithm Details

The Facial Emotion Recognition system uses the FER library with mtcnn=True enabled to guarantee reliable face detection in video frames.

1. **Video Processing:** The system gets a video file and processes it frame by frame.
2. **Face Detection:** For each frame, the emotiondetector.detect_emotions (frame) function is applied to locate faces.
3. **Emotion Classification:** Using an FER model, typically skilled on RGB images, the noticed faces are classified into emotions such as *sad*, *angry*, or *neutral*, with corresponding confidence scores.
4. **Aggregation:** After examination, the system defines the most assured prediction per detection and tallies the frequency of each primary emotion (e.g., 38 *sad*, 130 *angry*, 10 *neutral* in one case).

Facial signals serve as crucial indicators of stress and emotional conditions.

It combines CNN(Convolutional Neural Network) for extracting spatial features from images (edges, textures, facial landmarks, etc.) and RNN (Recurrent Neural Network) for video FER → captures how expressions change over time (e.g., a smile forming, eyebrows raising).

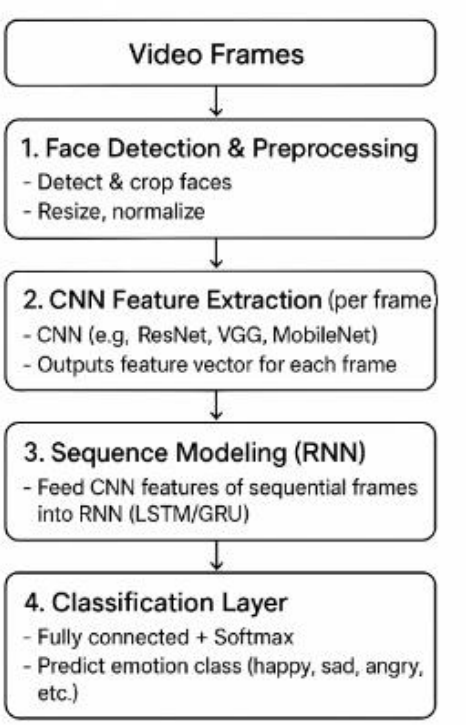

Figure 3:CNN & RNN for FER

3.6 Fusion Algorithm details

Figure 4 illustrates the basic structure of the multimodal system. In multimodality, the process of combining data from two or more modalities to perform prediction, classification, or regression is referred to as fusion. The multimodal data fusion approach is defined by the stage in the network where the input modalities are merged, which can occur at early, intermediate, or late levels.

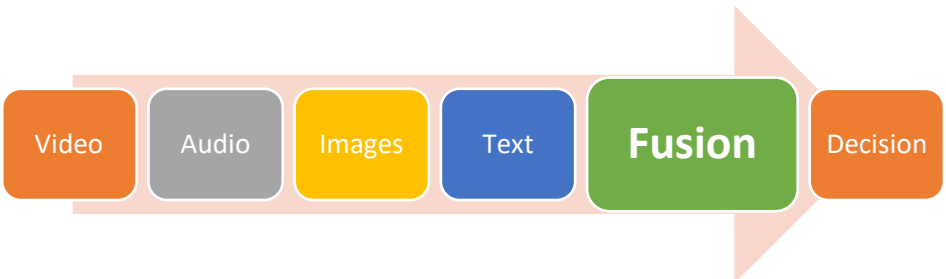

Figure-4

4. Experimental Setup

The system is developed and assessed within a dedicated video conferencing analysis setting, while also incorporating broader experimental insights from related studies.

4.1 Data Sources

4.1.1 Video Conference Data (Primary Notebook Focus)

The primary input for facial emotion recognition and audio transcription is a video file.

4.1.2 Microsoft FER +

FER+ is an updated and extended version of the original FER2013 dataset. With refined labeling, it offers more reliable ground truth for building and evaluating facial emotion recognition models.

4.2 Model Configuration

4.2.1 FER Model

For face detection and alignment, the Facial Emotion Recognition (FER) library is initialized with mtcnn=True, enabling the use of Multi-task Cascaded Convolutional Networks (MTCNN).

4.2.2 Whisper Model

The system employs OpenAI's Whisper model, specifically the "large" variant, for Automatic Speech Recognition (ASR). Whisper is a Convolutional Transformer-based end-to-end ASR model. Other versions, such as "base," "medium," and "large-v2," have also been utilized for tasks like language diarization and fine-tuning.

4.2.3 Sentiment Analysis Algorithms

NLTK VADER Sentiment Analysis:

Uses *nltk.sentiment.vader.SentimentIntensityAnalyzer* to generate polarity scores, including *neg*, *neu*, *pos*, and *compound*.

4.2.4 Tone Analysis Algorithm-Radial Basis Function Support Vector Machine (RBF SVM)

The Radial Basis Function Support Vector Machine (RBF SVM) is a robust machine learning algorithm suitable for both classification and regression tasks. As a non-parametric model, it is particularly effective when dealing with non-linear and high-dimensional data.

5. Results

The system produces a comprehensive assessment of employee feeling states through the analysis of multiple modalities. The results are quantified using various performance metrics.

5.1 Facial Emotion Recognition Results

From video analysis, the 'FER' library provides counts of detected primary emotions and their confidence scores

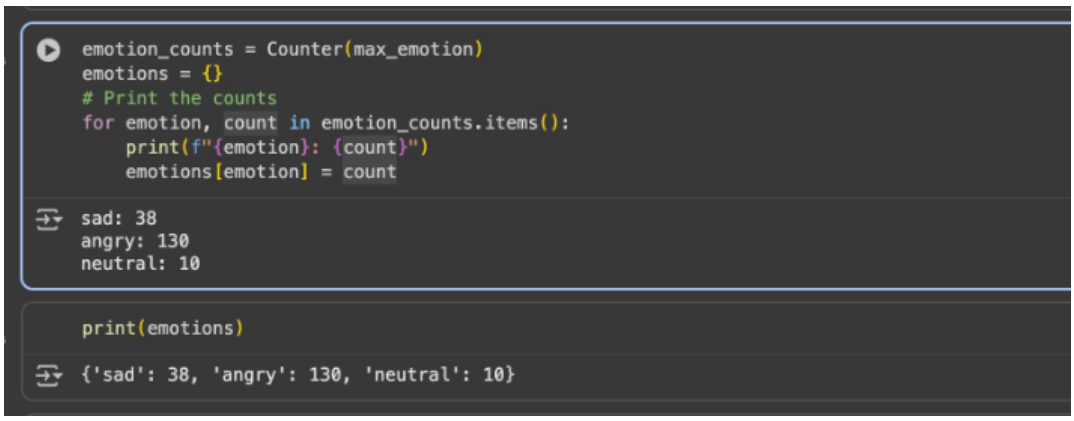

Figure-5

Above result shows occurrences like 38 'sad', 130 'angry', and 10 'neutral' detections, offering a direct measure of expressed emotions.It shows the person is angry

5.2 Speech Transcription Results(Text Modality)

The Whisper "large" model demonstrates optimal performance in Automatic Speech Recognition.

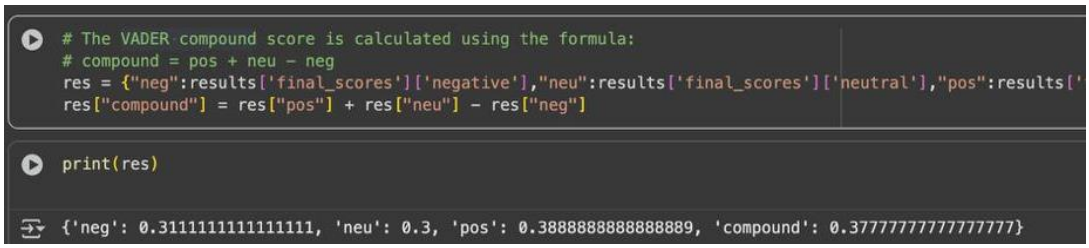

Figure-6

Above results shows tone is positive

5.3 Text Sentiment Analysis Results

• BERT-based Emotion Classification: The model, fine-tuned for 5-class emotion classification (Joy, Sadness, Neutral, Anger, Fear), achieved an accuracy of 94%

• NLTK VADER Sentiment Scores: The analysis provides polarity scores ('neg', 'neu', 'pos','compound'), indicating the overall positive, negative, or neutral tone of the transcribed speech

Figure-7

While the score is a continuous value, it's often used with a set of thresholds to classify the sentiment into categories:

Positive: Compound score ≥0.05

Neutral: Compound score between −0.05 and 0.05

Negative: Compound score ≤−0.05

The results shows compound score= -0.128. Which is Negative.

Negative emotion represents hostility, frustration, or antagonism (negative valence) or it represents a threat or danger, which is a highly unpleasant state (negative valence).

5.4 Multimodal Fusion and Overall Stress Detection Results

The result from Image sentiment analysis,text sentiment analysis and audio sentiment analysis are combined using Multimodal Fusion

```
# Combine the results
combined_result = combine_multimodal_sentiment(
    image_sentiment=IMAGE_SENTIMENT,
    text_sentiment=TEXT_SENTIMENT[0]['sentiment'], # Access the first element of the list
    audio_sentiment=AUDIO_SENTIMENT
)

print(json.dumps(combined_result, indent=4))
```

```
{
    "combined_sentiment": "Negative",
    "compound_score": -0.4327,
    "modality_weights_used": {
        "image": 0.25,
        "text": 0.4,
        "audio": 0.35
    }
}
```

Figure-8

Above results in Figure shows that Compound score=-0.04327 which is highly negative and we can conclude that the person in the input video is highly stressed.

Modal	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Image	72%	0.73	0.68	0.70
Text	76%	0.71	0.72	0.75
Audio	78%	0.76	0.74	0.75
Fusion	**88%**	**0.87**	**0.86**	**0.87**

Table 1-Stress detection accuracy, precision, recall and F1-score using single- and multimodality data.

5. Conclusion

In this work, a multimodal system for real-time detection of employees' emotional states during video conferences, with a focus on stress. The framework integrates Facial Emotion Recognition (FER), speech transcription using OpenAI's Whisper, and text sentiment analysis with NLTK VADER. The results indicate that fusing multimodal information for stress detection achieves an accuracy of 88%, offering valuable insights for future research on

multimodal stress detection using deep learning techniques. This study offers an objective framework for integrating multiple modalities through deep learning to detect stress, aiming to safeguard individuals from its adverse effects on physical and mental health.

References

[1] MS.N.Pavani, P.Supriya, A.SiriChandana3, B.Trinetra4, S.V.N.S.S.Supriya, "STRESS DETECTIONUSING IMAGE PROCESSING AND MACHINE LEARNING", Dogo Rangsang Research Journal, Vol-08 Issue-14 No. 02: 2021

[2] Jing Zhang et.al, "Real time mental stress detection using multimodality expressions with deep learningframework",frontiers in neuroscience,doi 10.3389/fnins.2022.947168

[3] Tanya Nijhawan, Girija Attigeri and T. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions",springer open access journal, https://doi.org/10.1186/s40537-022-00575-6,2022

[4] U SRINIVASULU REDDY, ADITYA VIVEK THOTA, A DHARUN, "Machine Learning Techniques for Stress Prediction in Working Employees", 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)

[5] D Y Liliana, "Emotion recognition from facial expression using deep convolutional neural network", 2018 International Conference of Computer and Informatics Engineering (IC2IE), Conf. Ser. 1193 012004

[6] Kusuma H R, Devika Rani B S, Anjali K, Ashwini N, "Stress Detection in It Professionals Using Image Processing and MachineLearning", International Journal of Advances in Engineering and Management (IJAEM), Volume 5, Issue 7 July 2023,pp. 255-259

[7] Dr. S. Vaikole, S. Mulajkar, A. More, P. Jayaswal, S. Dhas, "Stress Detection through Speech Analysis using Machine Learning", IJCRT | Volume 8, Issue 5 May 2020,pp. 2239-2244

[8] B. Padmaja, V. V. Rama Prasad and K. V. N. Sunitha, "A Machine Learning Approach for Stress Detection using a Wireless Physical Activity Tracker", International Journal of Machine Learning and Computing, Vol. 8, No. 1, February 2018, pp. 33-38

[9] Adnan Ghaderi, Javad Frounchi, Alireza Farnam, "Machine Learning-based Signal Processing Using Physiological Signals for Stress Detection", 22nd Iranian Conference on Biomedical Engineering(ICBME 2015), Iranian Research Organization for Science and Technology (IROST), Tehran, Iran, 25-27 November 2015

[10] Virginia Sandulescu, Sally Andrews, Nicola Bellotto, Oscar Martinez Mozos, "Stress Detection Using Wearable Physiological Sensors", International Work-Conference on the Interplay Between Natural and Artificial Computation,June 2015, DOI:10.1007/978-3-319-18914-7_55

[11] Xingxing Zhang , Chao Xu 1, Wanli Xue 1, Jing Hu , Yongchuan He and Mengxin Gao, "Emotion Recognition Based on Multichannel Physiological Signals with

Comprehensivenonlinear Processing", www.mdpi.com/journal/sensors, Sensors 2018, 18, 3886; doi:10.3390/s18113886

[12] Dorota Kami´nska, "Recognition of Human Mental Stress Using Machine Learning: A Case Study on Refugees", *Electronics* 2023, *12*, 3468.

[13] Reshma Radheshamjee Baheti, Supriya Kinariwala, "Detection and Analysis of Stress using Machine Learning Techniques", International Journal of Engineering and Advanced Technology (IJEAT), Volume-9 Issue-1, October 2019

[14] Manjunath R., Shivashankar, Shivakumar Swamy N, Erappa G, Manohar Koli, Nandeeswar S. B., Niranjan R. Chougala, "A Smart Biomedical Healthcare System to Detect Stress using Internet of Medical Things, Machine Learning and Artificial Intelligence", International Journal of Intelligent Systems and Applications in Engineering,Vol 11,Issue 4,2023,pp. 335-343

[15] Rohini Hanchate, Harshal Narute, Siddharam Shavage, Karan Tiwari, "Stress Detection Using Machine Learning", International Journal of Science and Healthcare Research, Vol. 8; Issue: 2; April-June 2023,pp-307-311

[16] Kavitha S Patil, Pranav Sivaprasad, Udhay Kiran K, Sujay G S, "Projective exploration on individual stress levels using machine learning", International Research Journal of Engineering and Technology (IRJET), Volume: 10 Issue: 04 | Apr 2023,pp-1449-1454

[17] Mohamed Razeed Mohamed Nowfeek, "A Review of Machine Learning Approach for Mental Stress Detection", Journal of Information Systems & Information Technology (JISIT), Vol. 6 No.2, 2021; pp- 72 – 83

[18] Nisha Raichur, Nidhi Lonakadi, Priyanka Mural, "Detection of Stress Using Image Processing and Machine Learning Techniques", International Journal of Engineering and Technology (IJET), Vol 9 No 3S July 2017

[19] Z. Zainudin, S. Hasan, S.M. Shamsuddin and S. Argawal, "Stress Detection using Machine Learning and Deep Learning", Asian Conference on Intelligent Computing and Data Sciences (ACIDS) 2021

[20] S. A. Singh, P. K. Gupta, M. Rajeshwari, and T. Janumala, ''Detection of stress using biosensors,'' *Mater. Today*, vol. 5, no. 10, pp. 21003–21010, 2018

[21] A. Alberdi, A. Aztiria, and A. Basarab, ''Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review,'' *J. Biomed. Informat.*, vol. 59, pp. 49–75, Feb. 2016.

[22] U. Pluntke, S. Gerke, A. Sridhar, J. Weiss, and B. Michel, ''Evaluation and classification of physical and psychological stress in firefighters using heart rate variability,'' in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 2207–2212

[23] G. Shanmugasundaram, S. Yazhini, E. Hemapratha, and S. Nithya, ‘‘A comprehensive review on stress detection techniques,’’ in *Proc. IEEE Int. Conf. Syst., Comput., Automat. Netw. (ICSCAN)*, Mar. 2019, pp. 1–6

[24] J. Wijsman, R. Vullers, S. Polito, C. Agell, J. Penders, and H. Hermens, ‘‘Towards ambulatory mental stress measurement from physiological parameters,’’ in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 564–569

[25] S. Elzeiny and M. Qaraqe, ‘‘Blueprint to workplace stress detection approaches,’’ in *Proc. Int. Conf. Comput. Appl. (ICCA)*, Aug. 2018, pp. 407–412.

[26] S. Elzeiny and M. Qaraqe, ‘‘Machine learning approaches to automatic stress detection: A review,’’ in *Proc. IEEE/ACS 15th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Oct. 2018, pp. 1–6.

[27] S. S. Panicker and P. Gayathri, ‘‘A survey of machine learning techniques in physiology based mental stress detection systems,’’ *Biocybern. Biomed. Eng.*, vol. 39, no. 2, pp. 444–469, Apr. 2019

[28] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, ‘‘Review on psychological stress detection using biosignals,’’ *IEEE Trans. Affect. Comput.*, early access, Jul. 9, 2019, doi: 0.1109/TAFFC.2019.2927337

[29] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, and C. Ersoy, ‘‘Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches,’’ *IEEE Access*, vol. 8, pp. 38146–38163, 2020

[30] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, ‘‘Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study,’’ *Sensors*, vol. 19, no. 8, p. 1849, Apr. 2019

[31] N. Keshan, P. V. Parimi, and I. Bichindaritz, ‘‘Machine learning for stress detection from ECG signals in automobile drivers,’’ in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Santa Clara, CA, USA, Oct. 2015, pp. 2661–2669

[32] V. Nasteski, ‘‘An overview of the supervised machine learning methods,’’ *Horizons*, vol. 4, pp. 51–62, Dec. 2017, doi: 10.20544/ horizons.b.04.1.17.p05.

[33] A. Cantara and A. Ceniza, ‘‘Stress sensor prototype: Determining the stress level in using a computer through validated self-made heart rate (HR) and galvanic skin response (GSR) sensors and fuzzy logic algorithm,’’ *Int. J. Eng. Res. Technol.*, vol. 5, no. 3, pp. 28–37, 2016

[34] G. Mattavelli et al, Facial expressions recognition and discrimination in Parkinson’s disease, J. Neuropsychol. 15 (1) (2021) 46–68

[35] Wan-Ting Chew et.al “Facial Expression Recognition Via Enhanced

Stress Convolution Neural Network for Stress Detection", IAENG International Journal of Computer Science, 49:3, IJCS_49_3_20, Volume 49, Issue 3: September 2022

[36] Menoua Keshishian et.al, "Understanding Adaptive, Multiscale Temporal

Integration In Deep Speech Recognition Systems", 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

[37] Terry Amorese et.al, "Automatic speech recognition (ASR) with Whisper:

Testing Performances in Different Languages.", S3C'23: Sustainable, Secure, and Smart Collaboration Workshop, Hosted by CHITALY 2023, SEPTEMBER 20–22, 2023, TURIN, ITALY

[38] Marian Pompiliu Cristescu et.al, "APPLYING BERT AND VADER IN HR SENTIMENT ANALYSIS", Journal of resource management and technology, CREATIVE SPACE ASSOCIATION | ISSN 2738-8719, Issue 2 / 2023

[39] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via largescale weak supervision," 2022. [Online]. Available: https: //arxiv.org/abs/2212.04356

[40] A. Siddhant, P. Jyothi, and S. Ganapathy, "Leveraging native language speech for accent identification using deep siamese networks," in 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2017, pp. 621–628.

[41] K. Kukk and T. Alumae, "Improving Language Identification of Accented Speech," in Proc. INTERSPEECH 2022, 2022, pp. 1288–1292.

[42] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPATDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in INTERSPEECH 2020, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.

[43] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva,F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021,arXiv:2106.04624.