Stress Detection at Workplace by Multimodal Analysis

Snigdha Mondal
School of Artificial Intelligence
and Data Sciences
Indian Institute of Technology
(IIT) Jodhpur, India
Email: mondal.3@iitj.ac.in

Arush Tripathi
School of Artificial Intelligence and
Data Sciences
Indian Institute of Technology
(IIT) Jodhpur, India
Email: tripathi.13@iitj.ac.in

***Abstract*—Stress is a significant societal issue, as it is the cause of many health problems and huge economic losses for companies. Detecting stress in computer users is technically challenging and, however, of the utmost importance in the workplace, especially now that remote working scenarios are becoming omnipresent. Owing to heightened competitiveness within the sector, companies are increasingly seeking enhanced efficiency and extended work hours from their personnel. However, with the existing deadline stress and all, they also face a work-life balance problem. To prevent stress from becoming chronic and provoking irreversible damage, it is necessary to detect it in its early stages.In this paper, we are proposing a multimodal system to detect the seven emotions of employees at the workplace. This study presents a novel COMBINED-STRESS model for detecting and analyzing stress. In our model, we try to imbibe this approach by fusing the 3 modalities, i.e., stress review data, audio, and face data, and predicting an output regarding the mental and stress health of the patient. In the audio model, we achieved an accuracy of 95.26% on training and 87% on the validation set using Bi-LSTM, whereas in the facial emotion detection model, we achieved an accuracy of 80% using the ViT and Bi-LSTM models combined. In the third model, we have used sentiment analysis and the PSS stress questionnaire to detect the stress percentage. At the end, we have combined all three models using the weighted combination of all three and predicted the final stress score/percentage of the employee.**

***Index Terms*—Stress, Bi-LSTM, ViT, Emotion detection, PSS test, MFCC, FER**

I. Introduction

Stress is a growing problem in our society. It is part of our daily life and many people suffer from it. The working conditions of professionals in India are deteriorating at an alarming rate in modern society. As per the Automatic Data Processing (ADP) Research Institute's most recent reports for 2021, 75% of Indian workers are stressed out at work, which is a major cause for concern [1]. Providing computer-based systems with the capability to recognize emotions is an ongoing subject of study. Consumer devices like, e.g., Laptops, computer, smartphone and home appliances should be capable to achieve an accurate reading of individuals' affective states, they could make appropriate decisions about how to interact with them, and adapt system's responses accordingly.

An Article by National Library of Medicine, [2] states that The Perceived Stress Scale is widely recognised as a prominent instrument utilised for the assessment of psychological stress. The self-reported questionnaire was developed with the intention of assessing the extent to which individuals perceive situations in their life as being stressful.Another report by Original Research article, [4] says that a temporal attention module (TAM),is capable of emphasising the distinctive temporal representation of facial expressions connected to stress using ResNet50 and I3D. On the other hand, [3] Voice Stress Analysis (VSA) offers a viable alternative for acquiring non-invasive means of extracting information regarding potential deceit from an individual's statement when they are experiencing psychological strain.Therefore, Extensive research in this particular domain has observed that individuals experiencing stress exhibit a range of complex indicators, which can be more effectively identified by a comprehensive examination of all three modalities.According to research findings [5], individuals experiencing depression frequently exhibit speech patterns characterised by stammering and irregular pauses. The patient also has a higher frequency of erroneous pronunciation. Through the utilisation of video modality, various additional aspects can be detected, such a typical eye contact, reduced frequency of mouth movement, altered posture, and so forth. By employing lexical analysis, it becomes possible to examine the linguistic context of the subject's speech, so yielding valuable insights on their mental well-being. By combining all of these channels, a comprehensive model may be constructed that incorporates all of these elements.Hence, We collected multimodality data from 110 participants during this research.The results demonstrate this COMBINED-STRESS model can accumulate stress-related information from multi-modality data to analyze a person's acute stress. It can serve as a tool for computer-aided stress detection.

II. Objective

The purpose is to identify a simple and practical method for assessing stress levels in individuals employed in the workforce. It is not always necessary for individuals in the working sector to utilise expensive equipment to measure stress levels, as these tools can be costly and may also subject individuals to peer pressure and the desire for recognition. Therefore, in order to address this issue, we propose a solution that is both cost-effective and reliable.

We have collected psychometric, facial and voice live dataset by generating google form so that we can accumulate their multimodal dataset to get more accurate result for Stress Detection. Specifically, we suggest employing a highly accurate model that has been developed based on our training data and Subsequently, the practise of incorporating real-time visual representations of the workforce has been identified as an effective approach. In our research, it has been determined that the ViT (Vision Transformer) with Bi-LSTM (Bidirectional Long Short-Term Memory) models exhibit superior performance in this regard.Hence, Our Image datasets have trained and evaluated in our experiment. Alternatively, In case of voice-based emotion detection Bi-LSTM has given the best result. Subsequently, following the administration of the Perceived Stress Scale (PSS) test, we have proceeded with the integration process and produce a conclusive report.

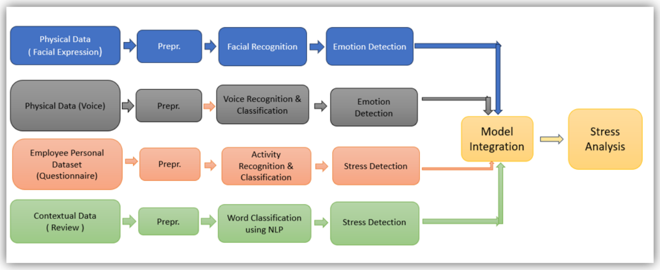

Fig. 1. Graphical Objective

III. Methodology

A. Approach Architecture

1) Personality/ Psychometric Test/ PSS Stress Analysis: There are four distinct approaches for the ultimate identification and analysis of stress. The initial method is the Perceived Scale Test(PSS), which stands as the most widely employed means of gauging an individual's stress level and capacity for tolerance. The Stress exam consists of a total of 20 questions, each offering 5 alternative options to pick from, The questionnaire has been designed based on the framework established by the World Health Organisation (WHO). Notably, there is a distinct feature in the questionnaire where 10 questions are given priority in descending order, while another set of 10 questions are given priority in ascending order.The Perceived Stress Scale (PSS) [10] is a well acknowledged and well-established instrument utilised for the evaluation of stress levels. The test indicated above, while being introduced in 1983, remains extensively employed to better our understanding of how different events affect our mental state and perceived levels of stress. Calculating the Perceived Stress Scale (PSS) Score The PSS score is computed by adhering to the prescribed instructions.

- To begin, it is necessary to invert the scores for questions and we set some threshold value for the mentioned questions and modify the scores for the option of the 10 questions as follows:

1= 5, 2 = 4, 3 = 3, 4 = 2, 5 = 1.
and
Never - 5, Rarely - 4, Sometimes - 3, Often - 2, Always - 1.

- Add the scores for each item and divide it by 100 to arrive at a total stress score

Individual PSS scores range between 0 and 1, with larger scores indicating greater perceived stress.

Scoring **below 0.33** would be termed **Low Stress**

Scoring range from **0 - 0.33** would be termed **Medium Low Stress**.

Scores range from **0.34 - 0.66** would indicate **Medium High Stress**.

Scores range from **0.67 - 1** indicate a significant level of **High stress**.

After collating result of all 20 Question :

$$\text{PSS Stress Score (pss)} = \frac{\text{Weightage of (Q1+Q2+}\ldots\text{+Q20)}}{\text{Sum of the Highest Weightage of Total Ques}} \tag{1}$$

2) Textual / Review based Sentiment Analysis: We have asked on the Google form for a review of their company's work culture; the response to this query will be used to conduct a sentiment analysis of the review, and a cumulative score will be provided based on the above stress test and this review sentiment score.

We have collected review from 111 Employee and Using TextBlob, Sentiment analysis has been performed by polarity range from -1 (Negative) to 1 (Positive) and got some compound score. Based on the Compound score a person's review is **Positive**, **Negative**, Or **Neutral**.

After that we need to normalize the Compound score as Negative reviews are giving negative(-ve) score so by normalizing this score it will help for further calculation as the value ranges from 0 to 1. Therefore suppose we get a high Normalized compound sentiment score, which indicates more towards Positive Review that implies Low Stress of a person.

Example:

Normalized Sentiment score = 0.75

Review Based Stress Score = 1 - Normalized Sentiment score = 0.25

a) Stress-point Textual Combined: Now, we have given weightage to the Psychometric Test and Review-based Sentiment analysis respectively 60% and 40% based on our believe in the genuity of the Dataset of a person.

$$SPS = 0.60*(PSSStressScore)+0.40*(ReviewBasedStressScore) \tag{2}$$

.

3) Facial Emotion Analysis: Background code will activate the webcam and begin detecting the user's facial expression when they begin filling out the form. It will capture a photo every 10 seconds; the minimum time required to fill out the form is approximately two minutes; therefore, there will be at least 12 to 13 photos of a person before the image is sent to an API call for emotion analysis. Following that, the employee's emotions and time signature will be saved to the database. We used the FER2013 Dataset to train the model with various facial expressions. In this data set, there are seven categories of facial expression: Angry, Disgust, Fear, Happy, Sad, and Surprise. Face images of 48x48 pixels are contained within the data. For facial emotion analysis we have used vgg16, Resnet50, Inceptionv3, Bi-Directional LSTM, Vision Transformer (ViT). By combining ViT + Bi-directional LSTM is giving 84% accuracy which is best out of all those model.

4) Voice-base Emotion Detection: Voice stress analysis (VSA) is a pseudoscientific technique that attempts to infer deceit from the measurement of tension in the voice. The technology seeks to distinguish between stressed and non-stressed outputs in response to stimuli (e.g., posed questions), with elevated stress considered an indicator of deception. We have collected voice for individual person through the google form by giving an option as uploading their voice. Here we have used MFCC (Mel-frequency cepstral coefficients) technique for feature extraction from audio signal as input to get better performance than directly considering raw audio signal as Input. We combined 4 Datasets: 1) RAVDESS dataset , 2) CREMA-D dataset , 3) TESS dataset, 4) SAVEE dataset to collect huge number of data which will help to prevent overfitting and getting more accuracy. After combing all those dataset we have used LSTM, Bi-Directional LSTM, Transformer. LSTM is giving 84.41% Tranning accuracy, 70% training accuracy for Transformer and 86.9% tranning accuracy for Bi-Dir LSTM. So we have choose Bi-Dir LSTM Model out of all those Model.

B. Model Generation and Architecture Application

Our proposed system involves several stages as shown in fig 2. First we collect the Psychometric, Textual and Voice-based data from employee using google form. During this time, the background code will turn on the camera and take a picture every 15 seconds until the g-form is submitted. Based on their responses, we set a threshold from 1 to 5 for each question and pre-processed their data using data normalisation to calculate their combined scores. Consequently, this stress score categorises an individual's emotional state into four distinct factors: High stress, Medium stress, Medium low stress, and Low stress. High and moderately high stress individuals will recommend medical consultation.

Thus, our methodology consists 3 steps : 1) Data Pre-Processing, 2) Feature Selection and 3) COMBINED-STRESS Model Integration

1) Data Pre-Processing:

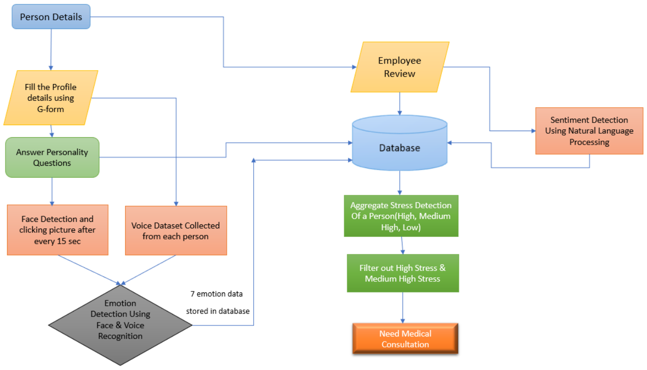
Person Details
Fill the Profile details using G-form
Answer Personality Questions
Face Detection and clicking picture after every 15 sec
Voice Dataset Collected from each person
Emotion Detection Using Face & Voice Recognition
7 emotion data stored in database
Employee Review
Database
Sentiment Detection Using Natural Language Processing
Aggregate Stress Detection Of a Person(High, Medium High, Low)
Filter out High Stress & Medium High Stress
Need Medical Consultation

Fig. 2. Proposed Model Flowchart

a) Textual/Review Based Data standardization: Dataset is preprossed and then fed into the Stress detection and Analysis Model. Data is collected in 2 ways through google form so in case of review based data analysis it's giving both positive and negative result. Hence, to combined this with PSS test result , we propose normalization to rescale the dataset result with in the range of 0 and 1. The formula for min-max normalization is:

$$Xnormalization = \frac{X - min(X)}{max(X) - min(X)} \tag{3}$$

Where X is a Review-based Data point and X normalization is a value between 0 and 1, we were able to calculate a combined Stress score from this and the PSS test.

b) Voice Emotion Based Data Refining: The following steps will be followed to preprocess the audio:

- Silence trimmed Sound by librosa.effects.trim()
- Defining the label in 3 classes : Positive (Happy) , Negative (Angry, Fearful, Sad) & Neutral (Calm , Neutral)
- Adding White Noise to make it robust to noisy and diverse environments
- random shifting to prevent overfitting by providing the model with a larger and more diverse training dataset.
- Pitch tunning that can imitate the natural pitch fluctuations caused by stress. This assists the model in identifying stress-related patterns across multiple intonation ranges.

c) Facial Emotion Based Data pre processing:

- Resize 48*48 Image into 224*224 resolution
- Convert RGB image into Gray Scale Image if color image is not required
- Normalisation pixel values to a specified range, such as [0, 1] or [-1, 1].
- Standardize pixel values using mean and standard deviation normalization

IV. Experimental Findings

This part provides a comprehensive analysis of the data obtained, encompassing a detailed examination of the proposed COMBINED-STRESS model. Our research seeks to address several key questions:

- **RQ1** Performance of Bi-LSTM Audio Modality with the used datasets?

- **RQ2** Performance of our Facial Modality using ViT+Bi-LSTM combination model?
- **RQ3** Working/Results of Textual Stress Test Model and Sentiment Analysis of Review Data?
- **RQ4** Working of COMBINED-STRESS model using all three modalities and Final model Score?

A. Research Settings

1) Dataset: This Project Model makes use of the three modalities; therefore, it uses three different types of datasets for its model to get kicked off. The project uses different datasets for Audio stress detection, Face Stress detection, and Stress detection through questionnaires and reviews of the inputs given by the employees. Let us discuss each modalities and their datasets in detail below:

a) Audio Modality: The Audio modality uses the combination of four datasets namely the RAVDESS dataset [7], CREMA-D dataset [6], TESS dataset [8], SAVEE dataset [9]. Each datasets mentioned above has certain characteristics that provide a much-needed boost to our audio speech model compared to its previous works, where they used only one or two of the above datasets. These 4 datasets are combined together to ensure that the model does not overfit.

b) Facial Modality: Another important modality used in this project is the facial data, this facial data taken from publicly available dataset FER-2013, The dataset comprises grayscale photographs of faces with dimensions of 48x48 pixels. The facial images have undergone automated registration to ensure that the face is approximately centred and occupies a consistent amount of space across all images.

c) Textual Modality: For this model, We have manually collected the data from the survey conducted and questionnaire prepared using the Perceived Stress Scale (PSS) assessment which is the global measure of conducting stress test, this test helps us understand how different situations affect ones feelings and his perceived stress. Every question consists of 5 points where 10 questions have different weightage and other set of 10 questions have different or reverse weightage from the first set. In this survey over 100+ working professionals participated and have given the input according to their work life condition. In addition to this , we have attached a separate question where an employee have to give the review about their work life balance of the employer and this review can be taken into account in the Sentiment analysis part of our model and based on the cumulative score the stress point can be calculated.

B. Evaluation

This section presents the experimental research evaluation of our COMBINED-STRESS model of audio, facial and textual modality of different models based on the research questions presented in the above section.

1) RQ1: Performance of Bi-LSTM Audio Modality with the used datasets: The first modality among all the three modalities used in the project is the Audio modality used for predicting the emotions and therfore is used to predict the amount of stress a person will be into by using the voice features. We have used these below features in our voice model to predict the emotions in our model-

- **Mel-Frequency Cepstral Coefficients**: captures the shape of the spectral envelope of a signal
- **Zero Crossing Rate**: captures the number of times a signal changes sign per second
- **Root Mean Square Energy**: captures the root mean square amplitude of the audio signal

Through these features we have extracted the emotions in the dataset. Also to make the model resistable to overfitting we have given our model four different types of data extracted from various sources , so that our model can learn and give the correct and unbiased decisions, and for this reason Our Bi-LSTM embedded audio model achieves the high training accuracy of 95.69% and validation accuracy of 87%.

Metrics	Accuracy	Precision	Recall	F1 Score
Weighted	0.869	0.869	0.864	0.865
Macro	0.864	0.8618	0.861	0.860

TABLE I
PERFORMANCE METRICS ON VALIDATION/TESTING SET

2) RQ2: Performance of our Facial Modality using ViT+Bi-LSTM combination model: In our second modality we have used the combination of Vision Transformer and Bidirectional-LSTM with the FER2013 dataset to train our facial modality. This achieves the validation accuracy of 80%. In our model we have used vision transformer because of following reasons:

- ViT Utilizes a transformer architecture that applies self-attention mechanisms to capture global relationships in the image.
- ViT Excels at capturing long-range dependencies and global context due to its self-attention mechanism. This can be advantageous for tasks that require understanding relationships between distant image regions.

The main reason to use vision transformer is to extract the features of the images , it is then sent to Bi-LSTM for sequential processing and for gathering the relationship of the features. ViT uses raw images and extract the patches of the image and the Bi-LSTM will then can capture temporal dependencies and relationships between the patches.

Our model achieves the following accuracy with this as shown in the below table :

Metric	Weighted	Macro
Precision	0.78	0.76
Recall	0.80	0.72
F1 Score	0.77	0.71
Accuracy	0.80	0.80

TABLE II
PERFORMANCE METRICS COMPARISON

3) RQ3: Working/Results of Textual Stress Test Model: Our third and Final Modality of the whole paper is the textual modality, the data of this modality is collected by conducting the survey of the working professionals from different companies, we have asked some questionnaire related to work culture based on the PSS test which is the universal test of stress calculation. Also , the employees are asked to provide the written review , so that the sentiment analysis can be performed in that. The combination of Stress test with sentiment data is taken into account for the final stress test. The min-max normalization is performed on the sentiment data , so that the nehative and positive review scores can be on the same page and we can then give the reading from 0 to 1. The calculated algorithm in our case to calculate stress using the Textual modality is given by:

$$Stress\_Score = 0.6*(Stress\_test\_Score)+0.4*(Sent\_Sc) \tag{4}$$

here sent_sc is sentiment score

We have given the weightage as 0.6 to stress test because the number of questions in test is more so the weightage is given to stress test more.

4) RQ4: Working of COMBINED-STRESS model using all three modalities: Our model was designed using the combination of the three modalities which are audio modality, facial modality and textual stress test modality. We have used three different modality because to get the more accurate prediction we can't rely on the one model's performance , the combined model can be helpful in deciding the true emotion of the people. Every person voice can be low or due to some health problem the shakiness might occur or his facial expression can be bad but may be he was not that unhappy or stressed that's why we have included one more modality which is stress test using questionnaire and review by the person himself. This can be essential as this will give the more accurate result along with the above models.

- **Audio Modality**: First we have used the Audio modality and obtained the accuracy of 95.26% in training set and accuracy of 87% in validation test set and evaluated our Audio Bi-LSTM model using other performance metrics such as macro recall, weighted recall, macro precision, weighted precision, macro F1-score, weighted F1-score, FPR and FNR.
- **Facial modality**: We have trained the facial modality and used it to predict the emotion of the person using the picture provided to our model. We have used ViT + Bi-LSTM model here, This model achieved the success rate of 80% which is a success as previous research paper have the achieved the accuracy of about 77%. we have also evaluated our ViT + Bi-LSTM model using other performance metrics such as macro recall, weighted recall, macro precision, weighted precision, macro F1-score, weighted F1-score, FPR and FNR.
- **Textual Stress Questionnaire modality**: We have prepared the questionnaire using the PSS stress test where 100+ people participated in the survey. The stress is calculated according to the weightage a question carries. The review of the employee is used in the sentiment analysis and using the min-max normalization theorem we have normalize its value. Further the stress test is given the weightage as 0.6 and review sentiment analysis is given the weightage 0.4. Using the above said metrics we have calculated the stress final score.

Let us see the working of our whole model:

- First we have asked the employee to participate in the survey which is basically the stress test designed using Perceived stress scale, where points are different for different questions.
- On the time of filling the form he has to open his camera module where his photo is being captured while filling the form, The photo captured will go to our trained facial module, where it will predict the emotion based on the expression, features of the image. here in the figure 3 it is clearly visible that the person is 88% happy.

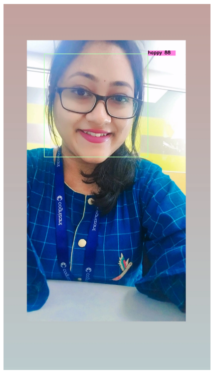

Fig. 3. facial detection model predicting emotion of the Employee

- Employee's review and stress score is calculated using equation (4) and the image model score is calculated.
- We should note that there are different score for different emotions predicted, the emotion marking table is below.

Predicted Emotion	Facial Stress Score
Happy	1 - (predicted score/100)
Sad	predicted score/100
angry	predicted score/100
disgust	predicted score/100
neutral	$\|(predictedscore/100 - 0.5)\|$
surprise	predicted score/100
fear	predicted score/100

TABLE III
EMOTIONS KEY

- In the above figure employee predicted score = 88 and emotion is happy , so by putting above table formula, the facial stress score = 0.12.
- Then at last employee's voice is taken into account, his voice score is calculated by putting his voice sample into our Audio modality, where the algorithm will predict what emotion the voice carries. The table for explaining emotion category for audio module is as follows:
- The below table will depict how we have assigned the weightage for each model:

Using all the above modalities, we have calculated the final stress of the employee score with the below algorithm.

Predicted Emotion	Stress Points	Verdict
Happy	0	Non-Stressed
Sad	1	Stressed
angry	1	Stressed
disgust	1	Stressed
neutral	0.5	Partial Stressed
surprise	1	Stressed
fear	1	Stressed

TABLE IV
AUDIO EMOTIONS KEY

Model	weightage
Audio	25%
Facial	30%
Textual	45%

TABLE V
WEIGHT DISTRIBUTION

FinalStressScoreCombined = 0.25*(Accuracy of Audio Modality)*Stressed/Not Stressed + 0.30*(Accuracy of Facial Modality)*Facial Stress Score + 0.45* Stress point Textual combined.

*where stressed is marked as 1, Partial stressed = 0.5 and non stressed as 0 in audio.

- The final Model Stress Score for some employees participated in survey are as follows:

Name	Audio Score(A)	Facial Score(B)	PSS Score(C)
Anusri	0.5	0.12	0.587
Gaurav	0.5	0.29	0.50
Prakhar	0	0.29	0.312
Kiledar	1	0.88	0.698
Vaishnav	1	0.63	0.472
Deependra	1	0.72	0.632

TABLE VI
MODEL SCORE OF SOME EMPLOYEES

Name	Final Str Score	Verdict	Remarks
Anusri	0.40185	Med. Low Str	-
Gaurav	0.40335	Med. Low Str	-
Prakhar	0.21	Low Stress	-
Kiledar	0.778	High Stress	Consultation
Vaishnav	0.5811	Med. High Str	Consultation
Deependra	0.6747	High Stress	Consultation

TABLE VII
FINAL MODEL SCORE OF SOME EMPLOYEES AND VERDICT

- Here low stress <0.33 , Medium Low stress <0.50 , Medium High stress <0.66 and High stress >0.66

V. CONCLUSION

The ”COMBINED-STRESS” model has been developed utilising three modalities. The first modality, known as the Audio Stress Model, employs a Bi-LSTM Architecture. This architecture has demonstrated a validation accuracy of 87%, surpassing the performance of many existing models. Furthermore, the training set accuracy stands at 95.26%. The second model employed in this study is the Facial Modality, which utilises a ViT+BiLSTM architecture to predict emotions based on facial traits. This model has demonstrated superior performance compared to various other prominent architectures, with an accuracy of 80% on the validation dataset. The final model, referred to as the Textual model, incorporates the PSS Stress test for measuring stress levels through the use of a questionnaire. Additionally, the model utilises NLP sentiment analysis of reviews to gain a more comprehensive understanding of the data. The three models were integrated and afterwards evaluated using a sample size of over 100 employees who participated in our survey. By applying our methodology, the stress score was ultimately derived. Subsequently, employees are requested to seek guidance from a medical professional or obtain a prescription based on their assigned Verdict category. This approach proves to be highly beneficial in many organisational and office settings, as it enables workplace proprietors to assess the well-being of their personnel. Consequently, this information can be utilised to strategically arrange suitable recreational and other activities, ultimately contributing to the organization’s financial gains.

ACKNOWLEDGMENT

The authors would like to thank Dr.Debasis Das, the project advisor, for affording the opportunity to partake in a research endeavour focused on the topic of Stress Detection. Additionally, we would like to extend the gratitude to our senior Ms. Bhumika, for her continuous support and guidance throughout this undertaking. Furthermore, we are thankful to our batchmates and coworkers for their significant assistance in collecting datasets during this venture.

REFERENCES

[1] S. D. Sharma, S. Sharma, R. Singh, A. Gehlot, and N. Priyadarshi, ”Deep Recurrent Neural Network Assisted Stress Detection System for Working Professionals,” in *Appl. Sci*, 2022.

[2] S. Meyer, G. J. Finn, S. E. Eyde, L. D. Kay, and K. L. Moreland, ”Psychological testing and psychological assessment: A review of evidence and issues,” *Am Psychol*, 2001.

[3] V. P. Patil, K. K. Nayak, and M. Saxena, ”Voice Stress Detection,” *International Journal of Electrical, Electronics and Computer Engineering*, 2013.

[4] J. Zhang, H. Yin, J. Zhang, G. Yang, and J. Qin, ”Real-time mental stress detection using multimodality expressions with a deep learning framework,” *Front. Neurosci*, vol. 16, 2022.

[5] P. Kaura, P. Jain, and M. Gupta, ”Depression Detection Through Multi-Modal Data - Final Report.”

[6] Haoyang Cao, David G. Cooper, Michael K. Keutmann, Raquel C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, 5(4), 377-390, October-December 2014. doi: 10.1109/TAFFC.2014.2336244. PMCID: PMC4313618.

[7] Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Funding Information: Natural Sciences and Engineering Research Council of Canada (2012-341583) doi: 10.5281/zenodo.1188976.

[8] M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto Emotional Speech Set (TESS). Borealis, 2020. Version: DRAFT VERSION. doi: 10.5683/SP2/E8H2MF. Available at: https://doi.org/10.5683/SP2/E8H2MF.

[9] Bogdan Vlasenko, Bjorn Schuller, Andreas Wendemuth, and Gerhard Rigoll. Combining frame and turn-level information for robust recognition of emotions within speech. In *Proceedings of Interspeech*, 2007, pp. 2249-2252.

[10] Cohen, S., Kamarck, T., Mermelstein, R. Perceived Stress Scale. *APA PsycTests*, 1983. doi: 10.1037/t02889-000. Available at: https://doi.org/10.1037/t02889-000.