Vision Transformer-Based Systems Achieve State- Of-The-Art In Stress Prediction From Facial Images

[1]ROSHINI JENIFER D
Computer Science & Engineering
Vels Institute of Science,
Technology and Advanced
Studies(VISTAS) Chennai, India
roshinicse02@gmail.com

[2] SHEELA GOWR P
Computer Science & Engineering
Vels Institute of Science,
Technology and Advanced
Studies(VISTAS) Chennai, India
sheela.se@vistas.ac.in

[3]THIRUMAL S
Computer Science & Engineering
Vels Institute of Science, Technology
and Advanced Studies(VISTAS)
Chennai, India
thirumal.se@vistas.ac.in

***Abstract*—Stress is the body's natural reaction to demands, which can cause tension in the body or mind. It is positive by helping people stay alert and motivated, but prolonged stress lead to negative health effects. The first step in the pipeline i s input data acquisition, where images often facial or biomedical in nature are gathered for analysis. These images undergo an initial preprocessing phase using CLAHE. This enhancement technique is both fast and lightweight, ensuring efficient processing while significantly improving local contrast, thereby making subtle differences in image regions more apparent. Next, the preprocessed images are sent through a segmentation module that utilizes a Wavelet Transform. This approach excels at retaining important image boundaries (edge preservation) and is robust to noise, yielding cleaner and more usable image segments for downstream analysis. The segmented images then undergo a feature extraction stage, utilizing Local Binary Patterns (LBP). LBP quickly summarizes local structures within the images, contributing both computational efficiency and the ability to discriminate subtle texture variations, an essential aspect in the accurate detection of stress indicators in visual data. Extracted features are subsequently fed into a classification module powered by a Vision Transformer. This modern architecture provides both high accuracy in classification tasks and interpret-ability, allowing the reason behind predictions to be better understood and trusted. The entire classification system is deployed using a Flask- based web application, integrating with web cameras for real-time data collection. This enables users to interact with the system and receive immediate feedback on their stress status. Finally, the user-facing interface makes the prediction output the outcome of this reliable pipeline is available, providing users with an evaluation of their stress level based on the analyzed images. This detailed process ensures accuracy, speed, and usability from gathering data to producing actionable prediction outcome**

Keywords—Stress Prediction, CLAHE, Wavelet Transform, Local Binary Patterns, Vision Transformer (ViT), Flask

I. INTRODUCTION

Stress is one of the most important problems of modern life, and it has become a contemporary living that affects people of all ages, occupations, and social backgrounds.

In today's fast-paced and competitive world, stress arises from a variety of factors such as work pressure and academic demands to financial problems and troubled interpersonal relationships, even to the disturbing constancy of technology. While temporary stress is occasionally a useful motivator and enhances performance in critical situations, long-term stress or chronic stress is highly destructive[1]. It weakens not only the body but also the mind, encouraging a variety of health disorders. Because of these risks, stress is now recognized not just as an emotional challenge but also as a major public health concern that requires early detection, proper monitoring, and effective management strategies[6].

The first step in recognizing the dual nature of stress is that it is not necessarily bad; often referred to as "eustress," or moderate levels of stress, can actually be adaptive in keeping people alert, focused, and motivated[7]. For instance, students studying for an exam or athletes competing in a tournament usually perform better under mild stress because this influence drives them to concentrate and channel their energy productively. If the level of stress becomes excessive, long, or overwhelming, it becomes harmful. This level, referred to as "distress," has the opposite effect; it lowers performance, clouds judgment, and adversely affects physical and mental health. Real-time interaction enhances usability because users receive immediate results of their stress levels [3].

In summary, stress is a normal aspect of human life, but unchecked stress can have serious risks to health and well-being. Traditional stress measuring techniques are either subjective or intrusive, highlighting the necessity for simple, non-invasive, and ongoing methods for stress recognition[2]. Body language, facial expression, and other behavioural cues provide useful information about stress levels and offer a practical way to identifying stress in real time situations.

II. RELATED WORKS

It reveals a strong trend toward combining physiological, behavioral, and computational methods to create comprehensive stress detection systems. Each approach has its own strengths and weaknesses, and research is ongoing aimed at balancing accuracy, non-invasiveness, and ease of use. The following sections provide a detailed overview of previous work in this area, focusing on the different methods, techniques, and results obtained by researchers.

Akhil Chandran Miniyadan et al [1] point out that a significant portion of the world's population suffers from depression and stress caused by various life factors. The fast-paced nature of modern society often cause stress, to build up and gradually develop into depression. The methodology involves using a Convolutional Neural Network to extract facial features from the FER dataset and incorporating the Patient Health Questionnaire. Experimental results show that the proposed system achieved a promising accuracy of 92% on the FER dataset. The main advantage of this approach is that visual and textual data can be integrated to improve prediction accuracy. However, a limitation is that the model

2026 Contemporary Computing Innovations Conference (CCIC) | 979-8-3195-2966-4/26/$31.00 ©2026 IEEE | DOI: 10.1109/CCIC68129.2026.11486015

performance may vary across populations due to bias in the dataset.

GioGiorgos Giannakakis et al [2] involves creating a dataset of 58 participants performing 11 stressful and non-stressful tasks across four experimental phases. Facial motion units are identified and extracted automatically as features. These characteristics are combined with conventional machine learning and deep learning models to analyze the effect of AUs through layer-by-layer association propagation. Advantages include high classification accuracy (>93%), identification of constraint-related AU combinations, and high interpretability of results. The system also supports custom modelling, improving the generalizability of data across individuals. However, drawbacks include dependence on controlled experimental settings, limited dataset size, and possible performance degradation in real-world settings. Additionally, relying solely on facial expressions may overlook other indicators of physiological stress.

Disha Sehgal et al [3] combine three modules to monitor mental health. First, we predict stress using behavioral and physiological data using the XG Boost algorithm[3], which is trained to effectively identify stress patterns. Emotions are then detected using a Convolutional Neural Network that classifies facial expressions and captures emotional information in real time. Third, a digital mood diaries allow users to record and track their emotional states over time. Key benefits include early detection of mental disorders, increased self- awareness, and availability of evidence-based mental health resources. However, drawbacks include potential data privacy issues, reliable on user compliance for diary entries, and possible misclassification due to model bias and limited datasets.

Sunitha C. K et al [4] describes this project involves several stages, starting with data preparation where clinical and biometric data are cleaned and normalized using Min-Max scaling. Multiple machine learning models such as decision trees, gradient boosting, linear regression, random forest, and SVM are trained using parameter hyper tuning to improve accuracy. To access stress in real-time, a script was developed to capture real data and provide predictions. Advantages include early stress detection, real-time monitoring, accessible graphical user-interface, and high predictive performance, especially with Gradient Boosting. However, disadvantages include dependence on high-quality clinical data, risk of model bias, and risk of over fitting with limited datasets.

Wenrui Dou et al [5] argue that as the proportion of the elderly population increases, facial expression recognition will become a major challenge for older adults as facial shape and texture change over time. In this study, we propose to use Vision Transformer network enhanced with LMIM to solve this problem. The LMIM method evaluates the correlation between the input image and high-level latent representations, allowing the model to identify the most important elements. This approach was evaluated on three datasets: EFED, a self-generated dataset for older adults, FER2013, and See pretty face, which served as a benchmark for performance evaluation. Experimental results shows that enabling LMIM improves recognition accuracy by about 4-9% compared to the baseline Vision Transformer. Overall, the combination of Vision Transformers with LMIM provides a promising solution for improving elderly facial expression recognition.

III. Proposed Methodology

A complete stress prediction system, however, requires more than just an improved classifier. The quality of the input data has a significant impact on the accuracy of predictions, especially when images are taken under different lighting and environmental conditions. To solve this problem, pre-processing using CLAHE is employed. CLAHE improves local contrast, highlights details, and provide reliable input data for analysis. Segmentation is then performed using a Wavelet Transform. This preserves important edges and reduces noise, and creates cleaner regions of the image for feature extraction. We then apply a Local Binary models to extract unique texture features, that capture micro-level changes in facial structures that correlate with stress[4]. These processed feature-rich images are fed to Vision Transformer for classification, achieving high performance and interpretability. The proposed pipeline is deployed through a Flask-based web application that supports real-time stress monitoring using web cameras as shown in Fig.1. This design ensures accessibility and providing instant feedback for users in various contexts such as workplaces, educational platforms, or medicial monitoring systems[12]. Integration lightweight pre-processing techniques with a powerful Vision Transformer classifier provides an effective balance between efficiency, accuracy, and real-world applicability.

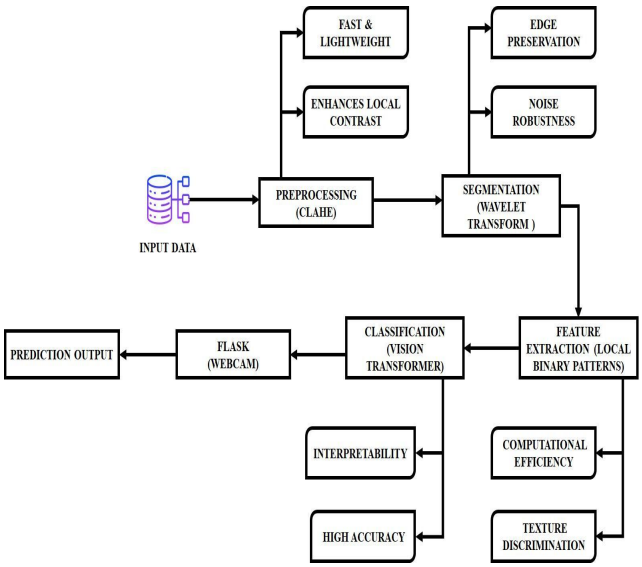

Fig. 1. System Architecture Diagram

A. Input Data

The pipeline starts with the input data. It primarily consists of facial images or biomedical images captured using cameras or other sensors. These images serve as the basis on which the entire stress prediction workflow is built. If the input images are poorly captured, blurry, or subject to lighting changes, the system may have difficulty detecting subtle features related to stress. Indicators of stress in facial images often include micro-expressions, changes in skin tone, wrinkles, or muscle tensions etc., and require clear and consistent input to be recognized accurately. Collecting different datasets ensures that the system is reliable for different people, age groups, and environmental conditions[2].

B. Preprocessing using CLAHE

Once the input data is acquired, it undergoes pre-processing using Contrast Limited Adaptive Histogram Equalization. Pre-processing is essential because real- world images often suffer from uneven lighting, low contrast, and environmental noise[13]. CLAHE is chosen because it is both fast and lightweight, meaning it requires minimal computational resources while being highly effective. It works by improving the local contrast in small regions of the image rather than adjusting the entire image globally.

C. Segmentation using Wavelet Transform

Following pre-processing, the images move into the segmentation stage, where Wavelet Transform is used. Segmentation is a crucial step because it isolates regions of interest from the background, focusing on areas that are most likely to contain stress-related features. The Wavelet Transform is particularly effective because it excels at edge preservation, which ensures that important structural details of the face are not lost. Stress- related cues such as the tightening of facial muscles or fine lines around the eyes may be subtle, and wavelet segmentation helps highlight these boundaries accurately[19].

D. Feature Extraction using Local Binary Patterns

Once the images are segmented, Local Binary Patterns are used for feature extraction. LBP is a commonly employed texture descriptor that represents the local features of an image by assessing each pixel against its neighboring pixels[8]. The technique operates by transforming texture data into binary formats, which are efficient for storage and analysis. One of the main advantages of LBP is its ability to discriminate micro-textures such as wrinkles, skin irregularities, or tension lines, all of which may correlate with stress levels [11]. By summarizing local structures, LBP reduces the complexity of the data while preserving the most important features.

E. Classification using Vision Transformer

The features obtained are subsequently input into a Vision Transformer (ViT) for categorization. Vision Transformers have recently become a leading approach for computer vision tasks because of their capacity to capture long range dependencies through self-attention mechanisms[5]. Unlike CNNs, which focus on local receptive fields, Vision Transformers divide the image into patches and analyse relationships across all patches simultaneously. For example, a Vision Transformer can correlate micro-level features like wrinkles with macro-level expressions such as furrowed brows or tightened lips [5]. One of the major advantages of ViTs is their interpretability, as attention maps can highlight which image regions were most important for the prediction.

F. Flask(Web Cam Integration)

For practical usability, the entire pipeline is deployed through a Flask-based web application. Flask is a lightweight Python framework that facilitates the connection between machine learning methods and user interfaces. Through this system, a webcam is connected to continuously capture real-time facial images. These images are immediately sent through the pipeline pre-processing, segmentation, feature extraction, and classification before generating predictions. Real-time interaction enhances usability, as users receive immediate feedback about their stress levels[14].The Flask interface also ensures scalability, making it possible to deploy the system in diverse environments such as workplaces, classrooms, or telemedicine platforms[16].

IV. Results and Discussion

A. Image Distribution for Train and Test Set

The distribution of images in the training and testing sets across the three stress categories: High Stress, Medium Stress, and Normal is shown in Fig.2. The training set is shown by the green bars, while the testing set is shown by the purple bars. Each class has a unique number of photos split between training and testing, to ensure that the representation of all categories in both subsets. The chart clearly shows how the dataset is divided for experimental purposes[10]. An overview of the data used to create and access the stress prediction system is shown in this visualization [15].

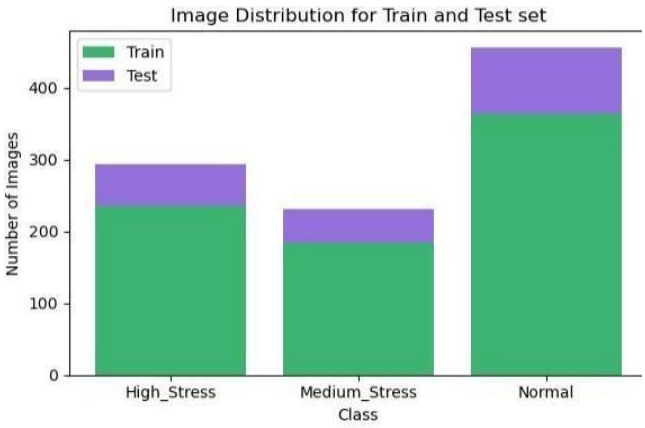

Fig. 2.Image Distribution for Train and Test Set

1. Random Train Images

Six grayscale pictures of faces with the labels "High Stress," "Medium Stress," or "Normal" are displayed in Fig.3. These images are likely part of a training dataset used for deep learning models that analyze facial expressions to classify stress levels. The top row shows two examples of high stress and one of medium stress, while the bottom row provides one medium stress example and two of normal emotion. This figure, captioned "Random Train Images," suggests a system is being trained to recognize and categorize different emotional states based on facial cues, which is a common application in fields like affective computing[18].

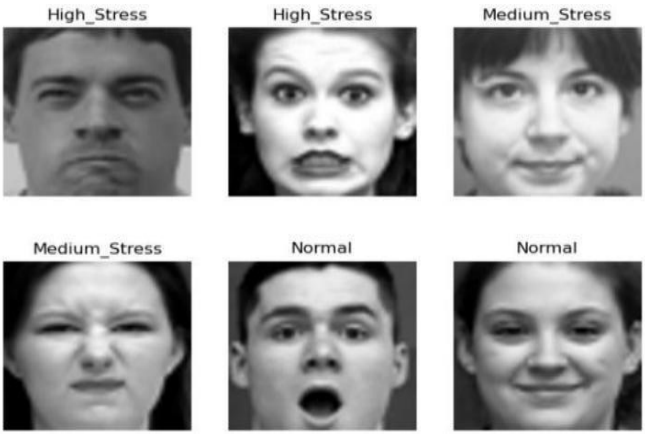

Fig. 3. Random images

2. Input Image

An "Input Image" is the raw image data as seen in Fig.4. That a machine learning model or an image processing algorithm is given to analyze. This image serves as the initial data point for a system to perform a specific task, such as classification or detection. In the context of the previous image of faces, the "Input Image" would be a new, unseen face that the model needs to classify as High Stress, Medium Stress, or Normal based on what it learned from its training data. The quality and characteristics of this input image are crucial for the model's performance.

Fig. 4. Input Image

B. *Resized RGB Image and Grayscale Image*

The image shows a side-by-side comparison of the same person's face as grayscale and resized RGB image. Due to its reduced quality, the resized RGB image on the left seems blocky and pixelated, revealing individual pixels. The grayscale image on the right has a better apparent resolution and looks smoother after being converted from a full color shades of gray. In order to reduce data quantity and computational complexity without sacrificing crucial features for tasks like facial recognition or emotion analysis[3], this comparison demonstrates a common steps in picture preprocessing for machine learning for converting photos to a standardized format and resolution as Fig. 5.

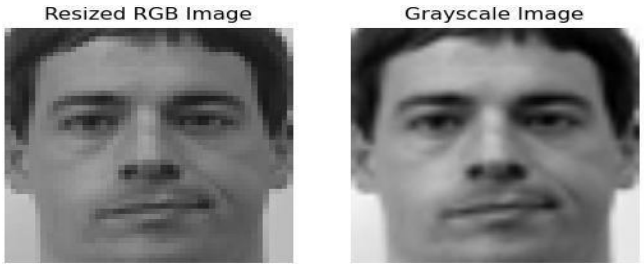

Fig. 5. Resized RGB Image & Grayscale Image

C. *CLAHE Image and Wavelet Transform Approximation*

This image is a visual comparison of two different image processing methods (CLAHE and wavelet approximation) applied to the same face image. The CLAHE image on the left has increased contrast, making facial features such as eyebrows, eyes, and mouth stand out more clearly. This technique is used to improve the local contrast of an image, especially in areas of varying brightness. The wavelet approximation image on the right, created using the wavelet transform, has a blurred or simplified appearance due to the removal of high frequency details. This process is often used to reduce noise or compress images [17]. Comparing these two images shows that the CLAHE image retains fine details while improving contrast, which is useful for feature extraction as Fig. 6.

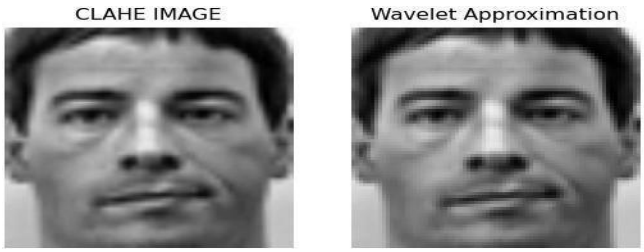

Fig. 6. CLAHE Image & Wavelet Approximation

D. *Local Binary Patterns*

The image displays the result of applying a Local Binary Pattern (LBP) algorithm to a facial image. This technique is used in computer vision for feature extraction and is highly effective for tasks like texture classification and facial analysis[20]. The LBP algorithm works by creating a binary pattern for each pixel by comparing the value of each pixel to the value of its neighbors. For example, if the value of an adjacent pixel is greater than or equal to the Centre pixel, it will be assigned a value 1, and if it is less than that, it will be assigned a value of 0. These binary values form a unique code that captures local texture and structural information. The resulting image, appears as a grainy abstract pattern of black, white, and gray pixels as shown in Fig.7.

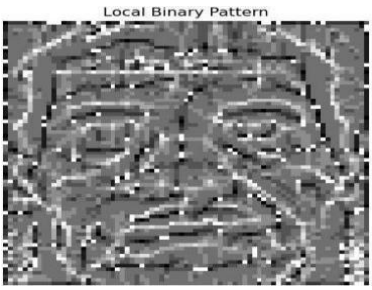

Fig.7. Local Binary Patterns

E. *Confusion Matrix*

The confusion matrix in Fig. 8 visually summarizes the performance of a machine learning model designed to classify stress levels. The rows of the matrix represent the true or actual stress levels of the images, and the columns represent the predicted stress levels determined by the model. For example, the cell where the "High Stress" row intersects the High Stress column will display the value of 57. This means that the model correctly classified 57 images as high stress. Similarly, the model correctly classified 45 images as Medium Stress and 92 images as Normal.

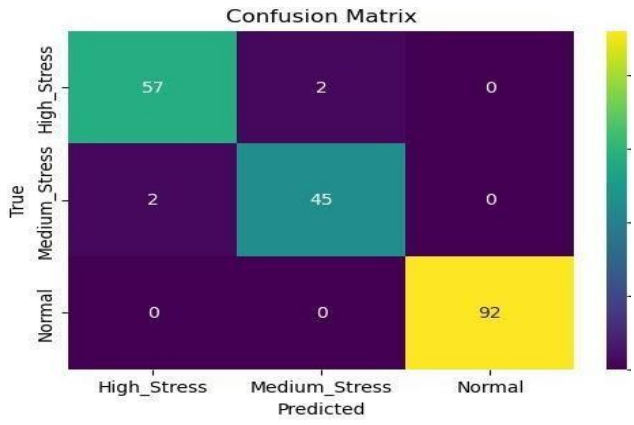

Fig. 8. Confusion Matrix

F. Validation Accuracy

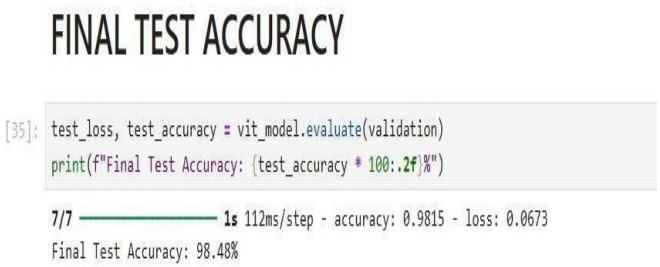

Fig. 9. Validation Accuracy

G. Classification Report

The image shows a detailed classification report for a deep learning model. Provides performance metrics such as precision, recall, and F1 score for each stress level: High Stress, Medium Stress, and Normal. This model shows high accuracy. It has a perfect score of 1.00 for "normal" face recognition and a high scores in the high 0.90 for all other classes. The overall accuracy of the model is 98%, indicating good performance. This report is an important tool for evaluating the reliability of the model in classifying different stress conditions as Fig.10.

Classification Report:

	precision	recall	f1-score	support
High_Stress	0.97	0.97	0.97	59
Medium_Stress	0.96	0.96	0.96	47
Normal	1.00	1.00	1.00	92
accuracy			0.98	198
macro avg	0.97	0.97	0.97	198
weighted avg	0.98	0.98	0.98	198

Fig. 10. Classification Report

H. ROC Curve

The image shows the ROC curve of a machine learning model, specifically Vision Transformer used to classify Stress levels. This kind plot tracks true positives against false alarms while shifting decision limits. Instead of staying flat, it climbs fast when accuracy improves. High performance means hugging the upper-left edge closely. The ideal curve would be along the top left corner of the graph. This figure has three separate ROC curves, one for each class: High Stress, Medium Stress and Normal. Each line lives near the top left area of the plot, just like in Fig. 11. That positioning suggests results close to ideal separation across all groups.

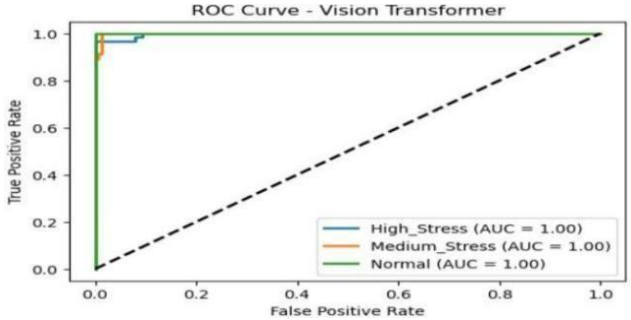

Fig.11. ROC Curve

I. Predicted Output

The provided image displays the result of a machine learning model's prediction on a facial image as Fig.12. The model has classified the person's expression as "Normal", with a confidence score of 0.5760. This score indicates that this model is 57.6% certain of its prediction[9].

Fig. 12. Predicted Output

The Fig. 13 shows the index page of the stress detection system implemented through a Flask-based web application. The interface provides a live webcam feed at the canter, enabling real-time image capture for stress prediction. Additionally, an option is available to upload an image and also user can instantly analyze their stress level by using the webcam, offering flexibility for users.

Fig. 13. Index Page (Live Webcam)

REFERENCES

[1] A.C. Miniyadan, P. P. G, N. G. P and R. R, 2025, "An Intelligent System for Prediction of Depression based on Facial Emotions and Textual Data," Emerging Technologies for Intelligent Systems, Trivandrum, India, 2025, pp. 1-6.

[2] Giorgos Giannakakis, Anastasios Roussos, Christina Andreou, Stefan Borgwardt, Alexandra I. Korda, 2025, "Stress recognition identifying relevant facial action units through explainable artificial intelligence and machine learning", Elsevier, Volume: 259

[3] D. Sehgal, D. Bansal, C. Singh and P. Jain, "Mental Health Awareness Using Machine Learning," 2025 International Conference on Networks and Cryptology (NETCRYPT), New Delhi, India, 2025, pp. 547-551

[4] S. C. K, R. S. A, D. M. V, S. B, S. V. K and M. C. K, 2024, "Synergetic Stress Insights: Unifying Facial Recognition and Sleep Patterns for Futuristic Stress Detection using AI," 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N), Greater Noida, India, 2024, pp. 1774-1779.

[5] W. Dou, K. Wang and T. Yamauchi, 2024 ,"Face Expression Recognition With Vision Transformer and Local Mutual Information Maximization," in IEEE Access, vol. 12, pp. 169263-169276,.

[6] C. Swedheetha, E. N. Sankar and C. B. Ramkishan, 2025,"Mental Stress Prediction Using Machine Learning and Facial Emotion Recognition," 2025 IEEE 14th International Conference on

Communication Systems and Network Technologies (CSNT), pp. 239-244.

[7] P. Gupta, S. Maji, V. K. Jain and S. Agarwal, 2024, "Automatic Stress Recognition Using FACS from Prominent Facial Regions," First International Conference on Pioneering Developments in Computer Science & Digital Technologies ,Delhi, India, 2024, pp. 488-492.

[8] K. Singh, S. K. Chawla, G. Singh and P. Soni, "Stress Detection using Machine Learning Techniques: A review," 2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2023, pp. 255-260.

[9] Chalapathi, Darshan, et al. "Biaxial deformation behaviour of duplex stainless steels: Experiments and crystal plasticity based stress predictions." Materials Science and Engineering

[10] Christ, Lukas, 2022 "The muse 2022 multimodal sentiment analysis challenge: humor, emotional reactions, and stress." Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge.

[11] S. R. Anthay, V. Nagarjuna, T. V. Mahesh, T. Aravind Royal and S.V.Varma, "Detection of Stress in Humans Wearing Face Masks using Machine Learning and Image Processing," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2022, pp. 1104-1110

[12] Shah, Milind, 2022, "Tool wear prediction in face milling of stainless steel using singular generative adversarial network and LSTM deep learning models" The International Journal of Advanced Manufacturing Technology, pp. 723-736

[13] Kumar Arora, Tarun, 2022, "Optimal facial feature based emotional recognition using deep learning algorithm", Computational Intelligence and Neuroscience.

[14] Mittal, Shivani, 2022, "How can machine learning be used in stress management: A systematic literature review of applications in workplaces and education." International Journal of Information Management Data Insights.

[15] Maurizi, Marco, Chao Gao, and Filippo Berto, 2022, "Predicting stress, strain and deformation fields in materials and structures with graph neural networks." Scientific reports

[16] Dalal, Surjeet, and Osamah Ibrahim Khalaf, 2021, "Prediction of occupation stress by implementing convolutional neural network techniques", Journal of Cases on Information Technology (JCIT), pp: 27-42.

[17] Stappen, Lukas, 2021, "The MuSe 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress." proceedings of the 2nd on multimodal sentiment analysis challenge, pp: 5-14

[18] Swami, Viren, George Horne, and Adrian Furnham, 2021 "COVID-19-related stress and anxiety are associated with negative body image in adults from the United Kingdom." Personality and individual differences

[19] Walambe, Rahee, 2021, "Retracted Employing Multimodal Machine Learning for Stress Detection." Journal of Healthcare Engineering.

[20] Jose Almeida and Fatima Rodrigues, 2021, “Facial Expression Recognition System for Stress Detection with Deep Learning”, 23rd International Conference on Enterprise Information Systems, - Volume 1, pages 256-26

.