Real-time Stress Detection through Facial Expressions Using a Vision Transformer

Aania Sohail[1], Mehreen Sirshar[2], Sumaira Shaukat[3]
Department of Software Engineering FJWU, Rawalpindi, Pakistan[1,2,3]
Email address: engr.aania.sohail@gmail.com (corresponding author)

***Abstract*— Stress is a common issue in today's society, impacting emotional health, decision-making, and work performance. This research introduces a real-time, non-contact method for detecting stress by analyzing facial expressions using deep learning. A Vision Transformer (ViT) model is fine-tuned to recognize facial features linked to stress and non-stress states. Unlike conventional techniques that rely on wearable sensors, this method uses a standard webcam, offering greater ease of use and accessibility. The model is trained on two publicly available datasets and assessed through accuracy, precision, recall, and loss metrics. It achieves an 86% accuracy rate in identifying stress in real time. This approach eliminates the need for physical devices, enhancing comfort and usability. It provides quick and automatic stress detection, making it suitable for use in practical settings such as schools, workplaces, and public areas. The system presents a scalable, user-friendly solution that supports emotional awareness and contributes to mental health monitoring.**

Keywords— Stress Detection, Facial Expressions, Vision Transformer, Deep Learning

I. INTRODUCTION

In today's era, stress has become an inevitable part of our lives. It is affecting individuals across many domains such as, education, personal relationships, and work. WinMR reported that in a recently conducted survey, 79% of people feel stressed and in 24% of those cases, work is the primary reason. Among university students, anxiety appears to be predominant; 88.4% of the students were found to be having symptoms of stress and anxiety as per NIH (National Institute of Health).

The stress that is caused due to various factors is invisible in nature but can be predicted by the human's behaviors and actions. The most natural form of non-verbal communication is the facial expressions, it gives an understanding of emotional states and stress is also considered an emotional state. Stress is a mixture of mental and physical reactions that take place when someone undergoes a difficult situation. When a person is in stress, it makes the person feel tensed or emotionally tired. The eyes might look wider in this state, the pupils get bigger and it feels like if the person is staring. In some situations, the eyebrows may rise, and some lines can be observed on the forehead, which shows that the person is emotionally upset. Similarly, lip compression and jaw tension are also the signs of stress. Like these symptoms, there are also some micro-symptoms that can be used to understand the small emotional reactions such as, slights change in the skin color and quick movements of facial features.

Over the years, researchers have explored different methods for stress detection, including psychological assessments and physiological signal monitoring using devices like ECG, EEG, or GSR sensors. These approaches provide valuable insights but they are usually expensive or impractical for real-time use in everyday settings. To address these issues, the field began shifting towards Artificial Intelligence AI-based solutions, utilizing the machine learning to recognize stress patterns from the behavioral data. As advancements in computer vision emerged, facial analysis became a prominent analysis offering real-time emotional recognition. The integration deep learning further enhanced it, by allowing the models to automatically learn complex facial features associated with stress. Convolutional Neural Networks (CNNs) were initially used for such tasks but this field is now moving towards more robust transformer-based architectures that capture both local and global patterns effectively. These developments have opened up new possibilities for real-time, accurate and scalable stress detection systems, particularly in environments like educational institutions and work places.

This research focuses on real-time stress detection through facial expressions using transformers. It is conducted in an educational institute of Pakistan. The Vision Transformer model is our primary deep-learning architecture for the stress detection.

II. RELATED WORK

The research involves the relevant literature on the problem we have worked on. An enhanced stress detection model was proposed (Chew, 2021). This model was designed using a deep learning-based Facial Expression Recognition dataset. They used a pre-trained 50-layer ResNet architecture to classify facial expressions related to stress. The model demonstrated improved accuracy by using convolutional neural networks for the extraction of features. This approach effectively identified stress-related emotions. This model was designed for real-world stress detection applications. A non-invasive stress detection method using face-related and emotion-related features extracted from facial video recordings was employed (Ogasawara, 2024). They utilized machine learning models to classify stress-related states, task recognition, and stress levels. Features such as Action units (AU), Face Embedding (FE), Valence and Arousal (VA), and Facial Emotion Recognition

were extracted and fused for better performance. They performed model training and evaluation. Their results showed that a fusion of Facial Expression Recognition dataset and AU (Action Units) achieved perfect stress state classification. This actually demonstrated the facial expression analysis for detecting stress. A systematic literature review (SLR) was conducted to analyse academic emotion classification using Facial Expression Recognition systems (Lek, 2024). They examined 701 publications across multiple databases, selecting 48 primary studies for in-depth analysis. The study identified deep learning approaches, especially convolutional neural networks (CNN), as the most widely used classifiers, with Facial Expression Recognition dataset and DAISEE dataset being the most commonly employed datasets. The review also highlighted the strengths and limitations of various feature extraction and classification techniques, noting that real-time expressions recognition systems remain less prevalent than non-real-time implementations. (Zhang, 2019) proposed a real-time stress detection framework based on facial expressions. Their approach classifies stress-related emotions, such as anger, fear, and sadness using convolutional neural network (CNN). The system first detects and aligns faces using Multi-task Cascaded Convolutional Networks (MTCNN), then extracts features from facial sequences and classifies them as stress-related expressions. If the number of stress-related frames surpasses a threshold, the framework issues a warning. The model was evaluated using CK+, Oulu-CASIA, and KMU-FED. This demonstrates real-world stress detection in a controlled environment. A stress recognition algorithm using general face images and facial landmarks to overcome the limitations of bio signal-based and thermal-imaging-based methods (Jeon, 2023). Their approach detects stress by analysing facial movements such as eye, mouth, and head position changes. They employed a deep neural network with shortcut mapping and bottleneck architecture to optimize the learning process and enhance recognition accuracy. The model was trained on a custom dataset with segmentation of three stress levels (no-stress, weak stress, strong stress), achieving the highest accuracy when using the grayscale images with facial landmarks. A stress detection system that classifies individuals as stressed or normal using physiological signals and Facial expressions (Sriramprakash, 2017), the study utilized stress-inducing tasks like email interruptions and time pressure conditions. Feature selection played a crucial role in improving classification accuracy. The results demonstrated that the combination of ECG signals and Facial Expressions can lead to accurate stress detection in working individuals. (Voleti, 2024) Proposed a real-time stress detection system using transfer learning techniques on facial expressions. The study compared two deep learning models, Mini-Xception and VGG16, to classify stress-related emotions. Facial features such as eyebrow movement, eye aperture, and lip compression were extracted from the FER dataset. Transfer learning was applied to enhance classification accuracy, with VGG16 achieving the best performance. The study demonstrated that deep learning-based facial analysis could effectively detect stress in real-world applications. A stress detection system using facial expression recognition powered by deep learning was F

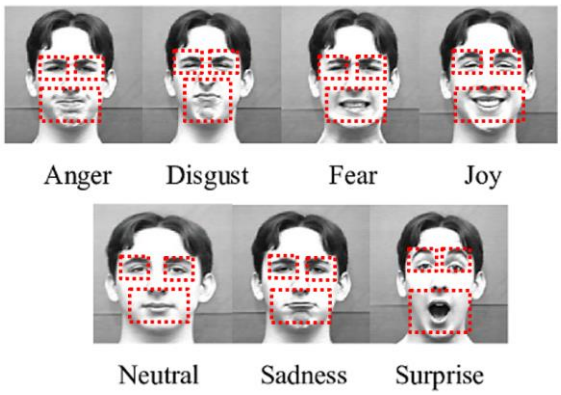


Fig 1: Seven Types of Facial Expressions

developed (Almeida and Rodrigues, 2021). They employed transfer learning with pre-trained models, including VGG16, VGG19, and inception-ResNet V2, to classify emotions from facial images. The system captures real-time images via a webcam and assesses stress levels based on detected expressions. The VGG16 model, combined with a convolutional layer-based classifier, achieved the highest accuracy in classifying stress-related emotions. The study highlights the potential of non-invasive, real-time stress detection using facial cues. (Zhang, 2022) proposed a real-time stress detection system combining ECG, voice, and facial expression analysis. The utilized ResNet50 and 13D with a Temporal Attention Module (TAM) to extract stress-related features from multimodal data. A matrix eigenvector-based fusion approach was employed to integrate these features for improved classification accuracy. The study used the Montreal Imaging Stress Task (MIST) to introduce stress in participants and collected multimodal data for validation. Their model achieved a moderate accuracy rate, demonstrating the effectiveness of deep learning-based multimodal fusion for a stress detection system. (Hills, 2019) investigated the impact of being observed on face recognition performance. They conducted three experiments to examine how social pressure and physiological stress influence recognition accuracy. Participants were observed during different stages of face recognition tasks, with stress measured using galvanic skin response (GSR) and heart rate. The study found that being observed during face learning reduced recognition accuracy, likely due to increased stress levels.

Fig 2: Stressed and No Stress Facial Expression

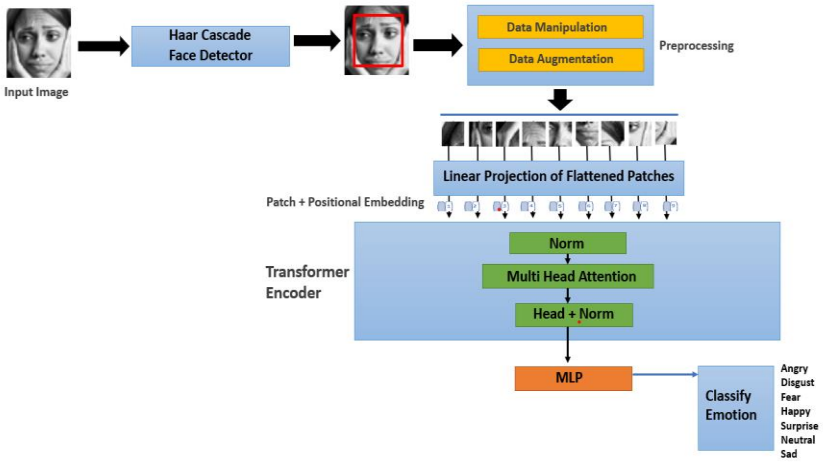


Fig. 3. Proposed architecture for Facial Expressions detection

The results highlighted the negative effects of social pressure on cognitive tasks. A real-time stress recognition system using wearable sensors and machine learning techniques. Their study used psychological signals such as heart rate and skin color to detect stress. A machine learning model was trained to process the sensor data, getting a moderate accuracy in detecting stress. This work was majorly completed for the wearable technology for continuous monitoring of stress in different situations (Jebelli, 2021). The study shows the implementation of pre-trained Convolutional Neural Networks utilizing AlexNet for facial emotion recognition (Shaees, 2021). The authors of this research have compared the performances of CNN models with traditional classifiers to highlight the effectiveness of deep learning approaches. (Oh, 2022) conducted a comparative analysis for emotion classification using the facial expressions data and physiological signals. Deep learning techniques were incorporated and multimodal inputs are combined to obtain improved accuracy of the system.

III. METHODOLOGY

This outlines the methodology used to achieve the objectives of the research.

A- Facial Expression Recognition

The transformer-based model architecture is followed for the facial expressions dataset to detect stress by mapping emotions to stress-related categories. The images from the dataset are first manipulated and augmented to enhance the readability and balance the samples. The grayscale images are divided into patches of a fixed size. The flattened patches are passed through linear projection to form patch embedding. Figure 3. Shows how the linear projection is applied on the image patches. The spatial order of the patches is preserved by applying positional encoding, aligning it in a sequence suitable for a Vision transformer model. The internal processing of the model includes Transformer Encoder Blocks which is implemented from scratch. Each block consists of layer normalization, multi-head self-attention, and a feed forward MLP layer. The output is processed via classification head which will predict the facial expression effectively.

B- Stress Detection

A pre-trained Vision Transformer model with a patch size of 16 is applied on the facial images from the stress-specific dataset. The images are then flattened and linearly projected into

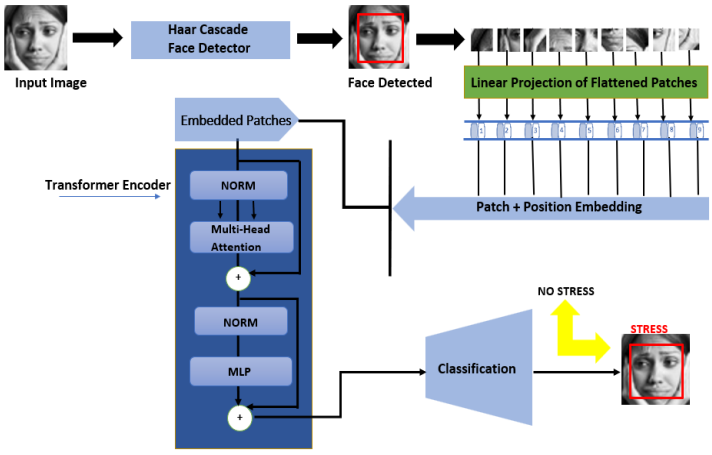


Fig. 4. Proposed architecture for the Stress and No Stress states Detection

embeddings. Positional encoding is applied to retain the spatial context before passing them into Vision Transfomer pipeline. Figure 4. shows the passing of image patches into the linera projection.

These patches are passed through the Transformer Encoder that will apply normalization, multi-head self-attention, and MLP layers to learn the comlex relationships across facial regions. Encoded representation us forwarded to the classification head which outputs one of the two labels: STRESS or NO STRESS. This is based on the learning of visual patterns from facial expressions in the dataset.

IV. EXPERIMENTAL SETUP

This section presents a detailed discussion on the experimental setup including the datasets descriptions, training strategies and evaluation metrics.

A- Datasets

The proposed framework is implemented on the two publicly available datasets

Facial Expressions Recognition:

The FER (Facial Expressions Recognition) dataset consists of seven classes of different facial expressions. These expressions are; Happy, Neutral, Surprise, Sad, Disgust, Angry, and Fear. Each class contains around 5000 images, and each image of 48x48 in size. This dataset comprises of grayscale images in which most of them are the frontal parts of the faces. This dataset involves the explicit classification of stress and no-stress expressions. The Happy, Neutral, and Surprise classes are classified as 'no-stress' facial expressions. Similarly, the other four classes, Sad, Fear, Angry, and Disgust, are marked as 'stress' classes.

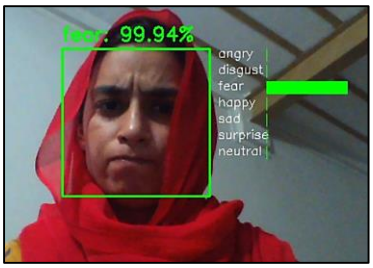


Fig. 5. Fear face expression detection

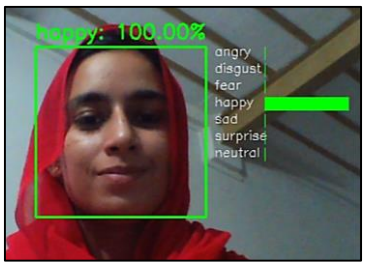


Fig. 6. Happy face expression detection

Stress Faces Dataset:

The second dataset being used is the Stress Faces dataset (SFD). This dataset consists of 2 classes 'stress' and 'no stress'. This dataset includes training and testing sections for each class. This helps in compiling the overall execution into one file for training the model. Both of the classes contain approximately 6000 images with a 48x48 size. The images in this dataset are grayscale images and frontal part of the faces. The images in this dataset show different lighting, head positions and facial expressions. Labels are based on actual signs of stress in the human body, like body or mind reactions.

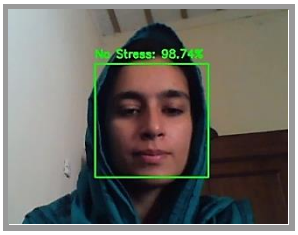
B- Training and Implementation Details for FER Dataset

The data for the model training is prepared by resizing and normalizing all the facial expression images to maintain the model consistency. The meaningful patterns from the visual data are extracted by the transformer-based deep learning approach. This model is structure with the aim to process the image features via multiple layers that are designed to effectively improve the learning efficiency while minimizing the chance of over fitting.

This training process is carried out of over several epochs. Optimization algorithm and loss function for multi-class classification are incorporated. During training, accuracy, loss, precision, and recall are continuously monitored. Techniques like early stopping and model check pointing are employed to make sure the obtained results are optimal as per the task at hand. The final evaluation includes both quantitative and visual analysis to measure the model's effectiveness.

C- Training and Implementation Details for the Stress Faces Dataset

The Facial images data is adjusted by resizing and normalization to improve rthe input quality. Data augmentation is carried out to improve the model's ability to generalize. A pretrained Vision Transformer models with a patch size of 16 is selected for image



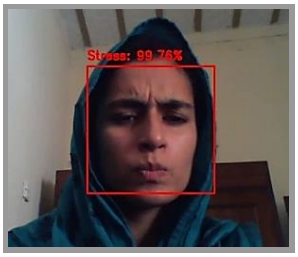


Fig. 7. Real-time execution through webcam

TABLE 1. EVALUATION METRICS OBTAINED FOR THE PROPOSED ARCHITECTURE

Proposed Methodology	Accuracy	Loss	Precision	Recall
Facial Expression Recognition	**99.1**	**0.09**	**0.99**	**0.99**
Stress state Detection	**86.7**	**0.3**	**0.87**	**0.86**

classification and it is fine-tuned specifically for stress detection. The original architecture is retained with minimal alternations to follow it for binary output. The training is implemented in a standard supervised learning approach. Evaluation metrics are observed throughout the training process to determine the learning behavor of the model. A continuous training setupt is used to make sure the reproducibility. Final evaluation is made via standard classification metrics, confirming the models' ability to differenciate between the given classes effectively.

D- Evaluation Metrics

Four essential metrics are considered to evaluate the performance of the model. Accuracy depicts the overall percentage of correct predictions. Precision shows the number of instances identified as stress that were actually correct. Recall is used to show the model's ability to detect all real stress cases. Loss shows the model's performance; lower loss means better performance.

V. RESULTS

This part of the research focuses on the results obtained after the model is trained and saved for the implementation.

A- Evaluation on FER Dataset

This execution is concluded on a training accuracy of 99.7%. It's considered an excellent approach in our case, as it leaves a few decimal points for the learning gap on the new data while running on the real-time webcam live data. The graphs obtained from the evaluation are visualized in the Figure 8.

The Confusion Matrix is generated to evaluate the classification performance of the trained model. Figure 9. This provides the information about model's ability to distinguish between all the classes.

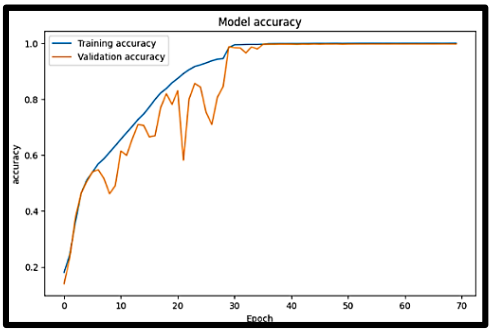


Fig. 8. Training and Validation accuracy graph for FER dataset

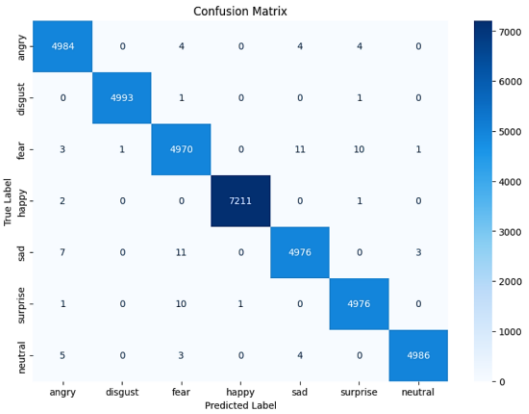


Fig. 9. Confusion Matrix for Facial Expressions Recognition Model

B- Evaluation on Stress Faces Dataset

The primary evaluation metrics used were accuracy, precision, recall, and loss. During training, we observed a consistent increase in accuracy and precision, with the training accuracy

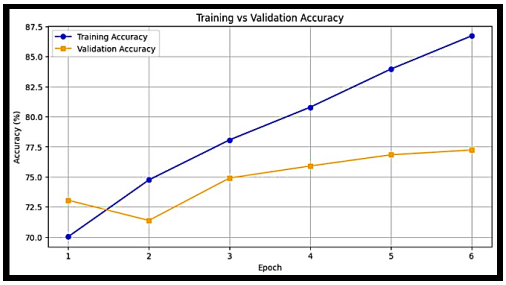


Fig. 10. Training and Validation accuracy graph for Stress Faces dataset

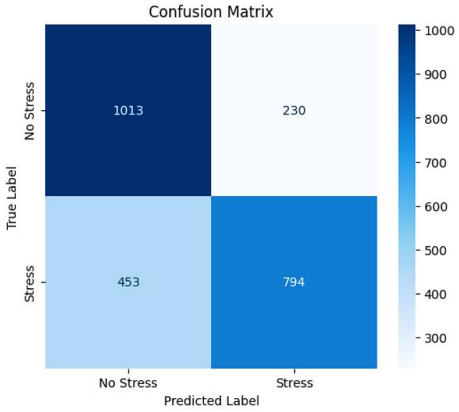


Fig. 11. Confusion Matrix for Stress Detection model

reaching 86.72% by the sixth epoch. Validation accuracy peaked at 77.23%, indicating the model was working. The best validation accuracy was obtained after the sixth epoch. Figure 10. The graphs obtained after the completion of training of the model are attached below.

The Confusion matrix (Figure 11) for the trained model to observe the class distinguishing ability is extracted for evaluation purposes. The matrix shows both accurate predictions and misclassifications across the dataset.

VI. CONCLUSION AND FUTURE WORK

The current system performs very effectively in real-time stress detection using facial expressions and a Vision Transformer model. This shows significant potential for future improvements. By merging some additional data sources like voice patterns, heart rate, or posture could make the model more accurate by providing a multimodal understanding of stress. Further training on larger and more diverse datasets would help improve the model's performance across various environmental conditions. To enhance the system's scalability and usability, optimizing the model for lightweight deployment on mobile or embedded devices is a key future direction. This would allow users to benefit from real-time stress monitoring without relying on high-performance hardware. Also, integrating a time-based analysis feature could help track user stress patterns over days or weeks, offering meaningful insights and early indicators of potential mental health issues.

This work presents a solid foundation for detecting stress through facial expressions using deep learning. The model achieved promising results in terms of accuracy and was successfully deployed for real-time detection. Although improvements are still possible, the system demonstrates how AI can be used to support mental well-being, and with continued development, it can evolve into a reliable, privacy-respecting tool for everyday use.

Acknowledgement

We would like to acknowledge the Fatima Jinnah Women University, Pakistan, for providing us with the resources in order to conduct this research.

Conflict of Interest Statement

All the authors declare that there are no competing interests that could influence the work presented in this article.

REFERENCE

[1] "Stress Levels rise around the world", WIN/Gallup International. (2022), https://winmr.com/stress-levels-rise-around-the-world/

[2] Part, C., Rosenblatt, J. D., Lee, Y., Pan, Z., lacobucci, M., Ong, A., & McIntyre, R. S. "The association between stress and severe mental illness: A meta-analysis". Psychiatry Research, 289, 113091, 2020

[3] Chew, W.T., Chong, S. C., Ong, T.S & Chong, L.Y. (2021). "Facial Expression Recognition via Enhanced Stress Convolution Neural Network for Stress Detection". VISAPP 2021.

[4] Ogasawara, R., Bouazizi, M., & Ohtsuki, T., "Camera-Based Stress Detection Using Face-Related and Emotion-Related Features". IEEE International Conference on E-health Networking, Application, and Services (HealthCom), 2024

[5] Lek, J. X. Y., & Teo, J., Academic Emotion Classification Using FER: A Systematic Review. Human Behaviour and Emerging Technologies, 2023, Article ID 9790005, 2024

[6] Zhang, J., Mei, X., Liu, H., Yuan, S., & Qian, T. (2019). Detecting Negative Emotional Stress Based on Facial Expression in Real Time. 2019 IEEE 4th International Conference on Signal and Image Processing.

[7] Jeon, T., Bae, H., Lee, Y., Jang, S. (2023). Stress Recognition Using Face Images and Facial Land Marks. Proceedings of the IEEE Conference

[8] Sriramprakash, S., Vadana, D. P., & Murthy, O. V. R. (2017). Stress Detection in Working People. Procedia Computer science 115, 359-366

[9] Voleti, S., NagaRaju, M. S., Kumar, O. V., & Prasanna, V. (2024). Stress Detection from Facial Expressions Using Transfer Learning Techniques. 2024 International Conference on Distributed Computing and Optimization Techniques (icdcot)

[10] Almeida, J., & Rodrigues, F. (2021). Facial Expression Recognition System for Stress Detection with Deep Learning. Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021), 256-263

[11] Zhang, J., Yin, H., Zhang, J., Yang, G., Qin, J., & He, L. (2022). Real-time Mental Stress Detection using Multimadality Expressions with a Deep Learning Framework. Frontiers in neuroscience, 16, 947168

[12] Hills, P. J., Dickinson, D., Daniels, L. M., Boobyer, C. A., & Burton, R. (2019). Being Observed Caused Physiological Stress Leading to Poorer Face Recognition. Acta Psychologica, 196, 118-128

[13] Jebelli, H., T., Kamat, V. R., & Ahn, C.R. (2021). A machine learning approach for real-time stress recognition using wearable sensors. IEEE Transactions on Automation Science and Engineering, 18(2)

[14] Shaees, S., Naeem, H., Arslan, M., Naeem, M. R., Ali, S. H., & Aldabbas, H. (2021, September). Facial emotion recognition using transfer learning. In 2021 International Conference on Computing and Information Technology (ICCIT-1441) (pp. 1-5). IEEE

[15] Oh, S., & Kim, D. K. (2022). Comparative analysis of emotion classification based on facial expression and physiological signals using deep learning. Applied Sciences, 12(3), 1286