Stress Detection Using Facial Emotion Recognition: Comparative Analysis of VGG16, ResNet-50, and a Custom CNN Architecture

1 Dr. S. P. Siddique Ibrahim
School of Computer Science and Engineering
VIT-AP University
Amaravati, India
siddique.ibrahim@vitap.ac.in

2 Praneetha Chaladi
School of Computer Science and Engineering
VIT-AP University
Amaravati, India
praneethachaladi@gmail.com

3 Bhavya Sri Galam
School of Computer Science and Engineering
VIT-AP University
Amaravati, India
galambhavyasri@gmail.com

4 Bharathi Varsha Maridu
School of Computer Science and Engineering
VIT-AP University
Amaravati, India
bharathi.maridu@gmail.com

5 Dhanalakshmi Matta
School of Computer Science and Engineering
VIT-AP University
Amaravati, India
mattadhanasree@gmail.com

***Abstract*—Stress is an important psychological factor that can significantly impact overall health and productivity; hence, timely detection and management of stress are essential to prevent adverse effects on mental and physical well-being. In this study, facial emotion recognition is employed to detect stress using the CK+ dataset, which is a widely used dataset for facial expression recognition, containing anger, disgust, contempt, fear, happiness, sadness, and surprise emotions. Initially, well-established architectures VGG16 and ResNet50 are tested on the CK+ dataset. Although these models delivered competitive accuracies, a custom hybrid model, based on Convolutional Neural Networks (CNN), achieved superior performance. The custom model used fewer parameters leading to faster training time while still maintaining a higher accuracy of 99%. The model outperformed already existing solutions in both accuracy and training time making it more effective for real-word stress detection through facial expressions. Notably, the custom model achieved comparatively higher accuracy than ResNet50 and faster training time compared to VGG16 model.**

Keywords—VGG16, ResNet50, Facial Emotion Recognition, Stress detection, Transfer Learning, CNN

I. Introduction

Stress is an often-cited psychological condition that has severe impacts on a human being's mental and physical health. It has been associated with a broad spectrum of health problems, such as cardiovascular diseases, depression, anxiety, and poor cognitive functioning.

There are many ways to identify stress, including physiological data and behavioural cues. Physiological approaches mainly rely on the tracking of HRV, EDA, and cortisol levels, as these are direct indicators of how the body reacts to stress. These techniques are very accurate but involve expensive equipment and can hardly be used in broad or real-time applications. On the contrary, facial emotion recognition has emerged as the alternative in detecting stress. This approach is based on sensitive facial information that can communicate emotional feelings through facial expressions and hence detect stress in a non-invasive and real-time manner. Furthermore, examination of emotions such as anger, fear, or sadness may also be utilized to measure the trend of stress through facial expressions.

In the present study, facial emotion recognition was selected as the primary approach to detect stress. With this in mind, for obtaining an integrated solution, transfer learning was used to take advantage of pre-trained models such as VGG16 and ResNet-50, which have been trained on large sets like ImageNet. Transfer learning is proven to be a good fit for applications where large datasets are not readily available or the task, such as stress detection, is a subtask of a more general task, like facial emotion detection The pre-trained models could differentiate between complex and simple features, such as edges and textures, that are crucial to efficient image recognition.

Fine-tuning pre-trained models on the relatively small CK+ dataset, specifically designed for the emotion recognition task, allowed the model to efficiently fit the particular task of stress detection without the need to manually start from scratch. This strategy would lead to faster convergence, a performance gain, and reduce overfitting, especially given the small size of the CK+ dataset. The benefit of transfer learning is substantial in this regard because it allows the deployment of pre-computed features already available and thereby boosts the precision and calculation speed of stress detection.

The latest developments in deep learning, especially those related to CNNs, have significantly improved the accuracy of facial emotion recognition, making CNNs an optimal choice for tasks related to detecting stress. By integrating transfer learning along with architectures of CNN, the study successfully built a custom model that eventually surpassed the superior accuracy level of detecting stress through facial expressions as compared to pre-trained models, thus highlighting the efficiency of this approach.

II. Related Work

This section outlines information related to some experiments, which have helped contribute to stress detection technology.[1] Stress detection is obtained using facial expressions by employing transfer learning using VGG 16 and Mini Xception models. VGG 16 outperformed Mini

Xception in stress detection accuracy with 97.5% and 70.5% respectively, and hence, is stated to detect small features of faces much better than Mini Xception as well.[2] The paper compares various classifiers used for the recognition of emotion leading to stress using MMI, JAFFE and CK+ dataset. The algorithms implemented are SVM, KNN, RF and MLPNN, PCA for dimensionality reduction and LBP, HOG and Gabor wavelets to extract features from the images. The highest accuracy is obtained by KNN with 93.46% CK+,followed closely by Random Forest which is 93.31%, SVM with 93.25%. [3] An algorithm that handles 1D ECG data and converts it into 2D one is proposed for stress detection using WESAD dataset. It also bypasses traditional time consuming feature extraction and filtering by utilising transfer learning. In addition, model compression utilizing quantization has reduced the computational footprint further achieving 90.62% accuracy. The method proved effective on low powered devices like mobile applications.[4] The paper presents a method of using the RNNs and Random Forest to detect mental stress based on EEG signals. The training and validation data comes from the GAMEEMO, SEED datasets. The classification is based on extracting Power Spectral Density (PSD) features. RNN got 87% accuracy for arousal and 83% for valence, whereas Random Forest got 83% for arousal and 75% for valence. [5] The paper uses Convolutional Neural Networks (CNNs) to detect stress based on facial expressions. The CNN algorithm is applied to the Kaggle face expression recognition dataset, containing about 71,000 images across seven. The proposed model achieved an accuracy of 85%, bypassing previous methods like KNN, which had 77.27% accuracy.

III. Proposed Work

A. Dataset Preparation

[8] This extended Cohn-Kanade dataset contains 593 video sequences derived from 210 subjects aged between 18 and 50 years. The population represented is 69% female and 81% Euro-American. Of the seven universal emotions proven through use of the Facial Action Coding System (FACS), such emotions include Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise. Each sequence represents an evolution from neutral to peak expression of emotion in 10 to 60 frames, at resolutions of 640x490 or 640x480 pixels. Furthermore, the dataset also contains 122 non-posed smiles, which make it dynamic facial recognition appropriate. Preprocessing of the CK+ dataset involves a step that prepares the dataset for training the model. After capturing frames from video sequences, the images are converted to grayscale. This reduces the computational complexity yet retains facial features needed in processing. The output images are resized to 48x48 pixels to standardize input dimensions for deep learning models. Normalization applies the pixel values by scaling the pixels to the range of [0,1]. This provides faster convergence for the model and stability in learning. There are also data augmentation methods used: "random rotations," flipping, zooming, and modification in brightness to artificially increase the size and variability of the dataset. Such techniques again reduce overfitting and enhance generalization capability of the model through simulation of various head tilts, orientations, lighting conditions, and facial positions. The emotions are classified into two classes for stress detection: stress and non-stress. The stress class contains anger, disgust, fear, and contempt, sadness which are usually collected under high stress conditions. The non stress class includes emotions like happiness, and surprise, indicating a relatively more relaxed emotional state.

B. Architecture

1. Convolutional Blocks

The Figure 1 represents the architecture of proposed model. The proposed architecture of the custom model is initiated with an input image of size (48, 48, 3), with four different convolutional blocks Across all convolutional blocks, batch normalization is applied right after each Conv2D to stabilize training. For each mini-batch, batch normalization calculates mean and variance of the feature maps output by the convolutional layer. Each activation is then normalized using:

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

Then, each feature is scaled and shifted using γ(scale) and β (shift) parameters using:

$$y_i = \gamma x_i + \beta$$

Batch normalization stabilizes the distributions of activations across layers and hence allows deeper networks to be trained effectively, significantly contributing to progress in image recognition and other applications. Followed by batch normalization, ReLU activation is applied to introduce non-linearity and, finally, dropout with the value of 0.25 to prevent overfitting. Initially, in the first block of two convolutional layers, 64 filters with a kernel size of (3, 3) and 'same' padding are applied. Following that, max pooling layer that reduces the spatial dimensions from (48, 48) to (24, 24, 3) is applied. The second convolutional block applies 128 filters with the same amount of padding but a larger kernel size, (5, 5). Further reduction of dimensions by another max pooling layer goes from (24, 24) to (12, 12). Finally, the third convolutional block uses 512 filters with a (3, 3) kernel size. Like the previous layers, it applies a max pooling layer that reduces dimensions once again to (6, 6). Finally, the fourth convolutional block applies another Conv2D layer with 512 filters with a kernel size of (3, 3), producing a feature map of size (3, 3, 512).

2. Flattening and Fully Connected Layers:

After the fourth Max Pooling layer, we obtain a feature map with dimensions (3, 3, 512). This feature map represents the spatial and depth related information extracted through four convolution blocks.

By using Flatten layer we transform the (3, 3, 512) feature map into a (1, 4608) feature vector. It allows the output of the convolutional layers (which are 3D) to be fed into fully connected (dense) layers, which operate on 1D vectors. We use a total of three fully connected layers. The first fully connected layer takes the (4608) feature vector and outputs a (256) feature vector that helps in reducing the complexity while maintaining the essential features extracted from the convolutional layers. Then the second fully connected layer uses batch normalization and ReLU activation to process the (256) vector and produces the final classification probabilities.). The final output layer consists of 2 units, representing the two classes: stress and non-stress . The output of these units undergoes a softmax activation function that provides a probability distribution over the two classes.

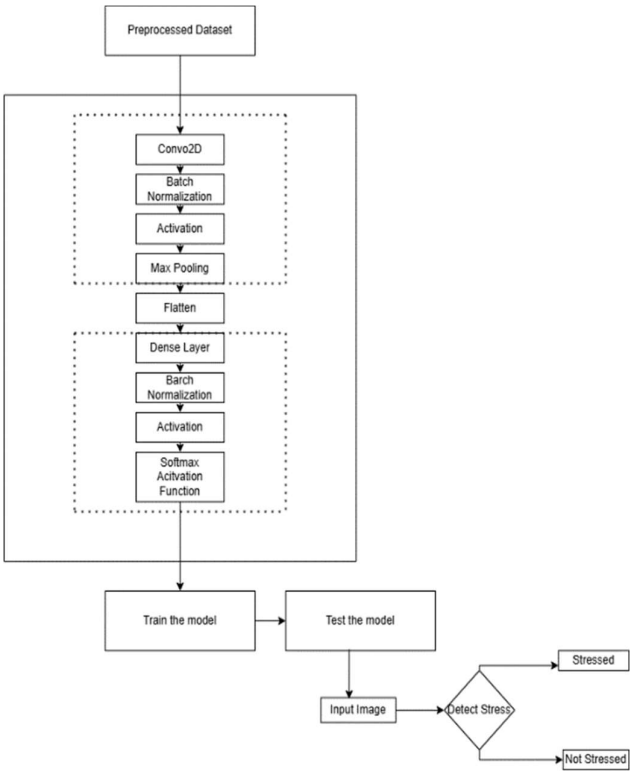

Fig. 1. Proposed model of the system

C. Methodology

This section portrays the methodologies applied for ResNet50, VGG16 and our proposed model for the stress detection using facial expressions.

1. ResNet50 and VGG16 Model

- **Step 1:** Import the pre-trained ResNet50 and VGG16 models. Exclude the top classification layer by setting include_top=False to fit to our particular task.
- **Step 2:** Freeze the base layers, so that the weights of the pre-trained layers do not update themselves during the initial training, and the model works upon the learned features.
- **Step 3:** Add a Global Average Pooling Layer such that spatial dimensions from base model are reduced to fixed-size feature vector.
- **Step 4:** Add a dense layer with 512 units using ReLU activation to capture more complex spatial pattern in the representation of features.
- **Step 5:** Introduce Dropout layer with dropout rate of 0.5 that reduces overfitting: dropping half of the neurons during training at random times.
- **Step 6:** Add the final Dense layer with a 2-neuron structure with softmax activation function for making binary classification of stress and non-stress states.
- **Step 7:** Use Adam optimizer, learning rate as 0.001 and categorical cross entropy loss function in the training of the model.

2.Proposed Model

This section describes the methodology applied for custom architecture in the facial expression-based stress detection model. The process mainly consists of three major components: **Image Preprocessing, Feature Extraction and Stress Detection**. Below are the steps followed to train and evaluate the model.

Component 1: Image Preprocessing

Input: CK+ dataset as facial images

- **Step 1**: Load the CK+ dataset, categorize the data into stress (anger, contempt, fear, disgust, sadness) and non-stress (happy, surprise) classes.
- **Step 2:** Normalize the pixel value to range [0,1], so that the networks converge faster during training.
- **Step 3**: Use data augmentation such as random rotation, flip, and zoom to increase variability in the dataset and thereby avoid overfitting.

Output: Preprocessed grayscale facial images

Component 2: Feature Extraction

Input: Preprocessed grayscale images

- **Step 1**: Pass images through four convolutional layers 64, 128, and 512 filters, and across all layers apply batch normalization, ReLU activation, max-pooling.
- **Step 2**: : This yields the final feature map of size (3, 3, 512).
  Step 3: : Flatten feature map to a 1D vector of 4608

Output: Each image has a flattened feature vector of size 4608.

Component 3: Stress Detection

Input: Flattened feature vectors.

- **Step 1**: Three fully connected layers are used for classification. The first dense layer will have 256 units, whereas the second and third will have 512 units. Add batch normalization, ReLU activation and dropout (0.25) across all layers.
- **Step 2**: The last output layer contains 2 units that represent the two classes: stress and non-stress.
- **Step 3**: Introduce the softmax function in order to include the likelihoods of the presence of either class; again, the assigned predicted label will depend on which is the larger of the two classes.
- **Step 4**: Compiler the model using Adam optimizer to minimize the loss function

- **Step 5**: The model then takes the test images and predicts some on the test data.
- **Step 6**: The performance obtained by the model is compared against the accuracy metric. The actual labels are used to compare predicted outputs of the model, and it calculates its accuracy to assess overall performance of the model.

Output: Predicted labels for stress or non-stress

Functions used:

1. Softmax Function: This is used in the output layer to convert raw model outputs to probability for each class.

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$

z_i is the output of class i
n is the total number of classes.

2. Optimizer function: The model is compiled using Adam optimizer. It uses learning rate of 0.001.

$$\theta_t = \theta_{t-1} - \alpha \cdot \frac{m_t}{\sqrt{v_t} + \epsilon}$$

where:

- θ represents the model parameters,
- α is the learning rate,
- m_t is the estimate of the first moment (mean),
- v_t is the estimate of the second moment (variance),
- ϵ is an constant to prevent division by zero.

3. Loss Function: For this model, categorical cross entropy is the loss function used, suitable for multi-class classification problems, and this computes the difference between the probability distribution that the model predicts and the real one.

IV. Results and Discussion

A. Discussion

The table below shows the accuracy achieved by each model when trained on CK+ Dataset:

Model	Accuracy
ResNet50	97.4%
VGG16	98.2%
Proposed Model	99.5%

1.Confusion Matrix:

The Fig.2 , Fig. 3, Fig. 4, represent confusion matrices of ResNet-50, VGG16, Proposed Architecture models respectively.

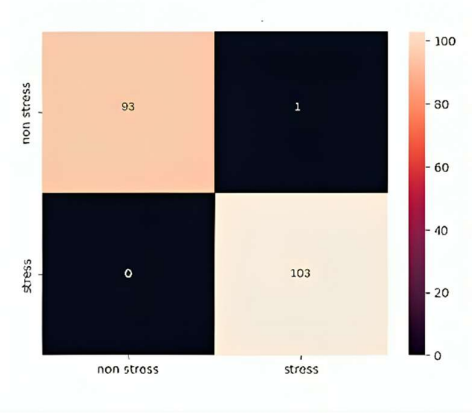

Fig. 2. Confusion matrix for Proposed model

True Positives:103
True Negatives:93
False Positives:1
False Negatives:0

The confusion matrix for the proposed custom model is capable of showing near-perfect performance in stress detection. It correctly classified 93 non-stress cases and 103 stress cases. Only 1 false positive occurs while there are zero false negatives (no case of a stress misclassified). It makes for very high precision and recall, with little misclassification. The model achieves a remarkable accuracy of 99.5% and generalizes very well, making an obvious distinction between stress and non-stress states. This low error value underlines the solidity of the customized model in real-world applications of stress detection.

Fig. 3. Confusion matrix for ResNet50 model

The matrix obtained in the confusion for this ResNet 50 model indicates good performance in detecting stress with a

relatively high accuracy of 97.46%. 93 stress cases were correctly classified, while 99 non-stress cases were also classified as True Negatives. Still, there were 4 false positives, where non-stress was indicated as stress, and 1 false negative, which is misclassified stress as a non-stress case. Although the model works well on the whole, 4 false positives and 1 false negative reveal space for improvement in classification precision and recall.

Fig. 4. Confusion matrix for VGG16 model

True Positives:95
True Negatives:99
False Positives:2
False Negatives:1

The confusion matrix for the VGG16 model reflects highly accurate performance in stress detection, with an impressive accuracy of 98.7%. The model correctly classified 95 non-stress cases and 99 stress cases, which points toward good generalization ability. However, there were 2 false positives, where non-stress cases were misclassified as stress, and 1 false negative, where a stress case was classified as non-stress. Despite these minor misclassifications, the model shows high precision and recall in distinguishing between stress and non-stress states.

2. Training and Validation Accuracy Curves

The Fig.5 , Fig. 6, Fig. 7, represent Model Accuracy Curves of ResNet-50, VGG16, Proposed Architecture models respectively.

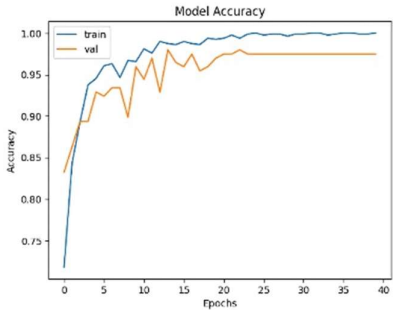

Fig. 5. Model Accuracy Curve for ResNet 50

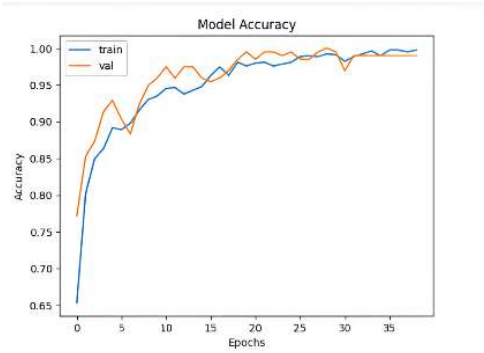

Fig. 6. Model Accuracy Curve for VGG16

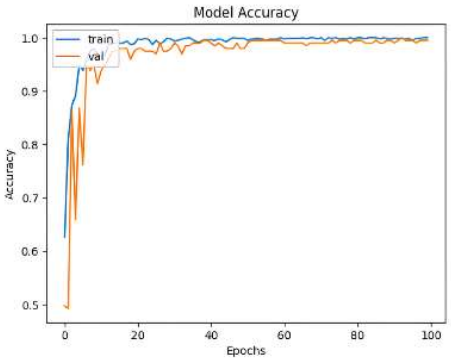

Fig. 7. Model Accuracy Curve for Proposed Method

Even though the gap between the training accuracy curves and the validation accuracy curves is not remarkable for both VGG16 and ResNet50, there could still issues regarding generalization on unseen data. Conversely, the proposed model is almost perfect; further minimizing the probability of overfitting whilst simultaneously increasing the capacity of the model's ability to be generalized further under a wider range of conditions. Additionally, there is an observance that proposed method was more accurate and stable with fewer numbers of epoch trained than VGG16 and ResNet-50 models, which had to use more epochs in order to obtain comparable levels of accuracy and stability.

3. Training and Accuracy Loss Curves

The Fig.8 , Fig. 9, Fig. 10, represent Model Loss Curves of ResNet-50, VGG16, Proposed Architecture models respectively.

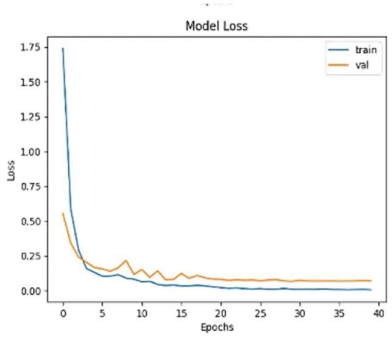

Fig. 8. Model Loss Curve for ResNet 50

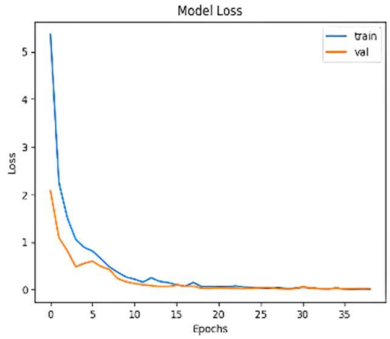

Fig. 9. Model Loss Curve for VVG16

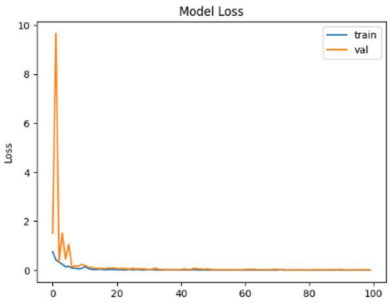

Fig .10. Model Loss Curve for Proposed Method

The gap between training and validation loss even though small, suggests that they might make classification errors when they face doubtful or complicated cases of stress and non-stress. In the Proposed Model, in both training and validation loss, exhibits a consistent and smooth curve with minimum divergence between these curves. This clarifies that the model is more optimized as there are fewer misclassifications and good generalization to new data.

The proposed model's architecture includes four convolutional blocks with batch normalization, ReLU activation, dropout, and max pooling, which ensures feature extraction while avoiding unnecessary complexity. Unlike deeper architectures of VGG16 that come with added parameters as well as skip connections in ResNet50, the proposed model is relatively streamlined that leads toward quicker training and lowers memory footprint. The design facilitates effective training since the model can figure out pertinent features quite fast without being overly complex, which contributes towards high efficiency performance. Architectural differences it grants for the proposed model—a streamlined design, proper dimensionality reduction as well as optimal regularization techniques contribute toward this ability of efficiently detecting stress from facial emotion recognition systems as compared to VGG16 and ResNet50.

V. Conclusion

In this study, established VGG16 and ResNet50 models and customized model for stress detection through facial emotion recognition are tested. The custom model achieved an accuracy level of 99.5% with few misclassifications proving its robustness and reliability for real-world applications in stress monitoring. Despite these strengths, there is scope for future development. Adding more data sources, including physiological metrics such as heart rate variability, is likely to further increase accuracy. Another way would be to explore techniques in transfer learning and hyperparameter optimization in order to improve the performance. An increased training dataset to include a wider range of facial expressions and diversity in demographics will also increase the adaptability across different populations. Refining the proposed model in these areas will increase its applicability and make a way for even more effective stress detection solutions in practical environments.

VI. References

[1] Voleti, Sravya, et al. "Stress Detection from Facial Expressions Using Transfer Learning Techniques." 2024 International Conference on Distributed Computing and Optimization Techniques *(ICDCOT)*. IEEE, 2024.

[2] P. Tiwari and S. Veenadhari, "An Efficient Classification Technique For Automatic Identification of Emotions Leading To Stress," *2022* IEEE 6th Conference on Information and Communication Technology *(CICT)*, Gwalior, India, 2022, pp. 1-5, doi: 10. 1109/ CICT56698. 2022.9997823.

[3] Ishaque, Syem, Naimul Khan, and Sri Krishnan. "Detecting stress through 2D ECG images using pretrained models, transfer learning and model compression techniques." Machine Learning with Applications 10 (2022): 100395..

[4] Khan, Md Raihan, and Mohiuddin Ahmad. "Mental Stress Detection from EEG Signals Using Comparative Analysis of Random Forest and Recurrent Neural Network." *2024* International Conference of Advances in Computing, Communication, Electrical, and Smart Systems *(iCACCESS)*. IEEE, 2024.

[5] M. Tarun, V. K. Jonnalagadda, A. J. Sai, K. Nivas and P. V. Mohan, "Stress Detection by Deep Learning Technique," *2024* Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing *(INCOS)*, Krishnankoil, Virudhunagar district, Tamil Nadu, India, 2024.

[6] Fasel, I., & Luettin, J. (2003). Automatic facial expression analysis: A survey. Pattern Recognition, 36(1), 259-275.

[7] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv : 1409.1556.

[8] S. L. Fernandes and G. J. Bala, "A Study on Face Recognition Under Facial Expression Variation and Occlusion," in Proceedings of the International Conference on Soft Computing Systems, 2016, pp. 371-377. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

[9] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 94-101.

[10] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 25, 1097-1105.

[11] Li, H., Zhang, Z., & Liu, L. (2019). Facial Expression Recognition with Convolutional Neural Networks via Transfer Learning. IEEE Access, 7, 1-12.

[12] A.N.Parab,D.V. Savla, J. P. Gala and K. Y. Kekre, "Stress and Emotion Analysis using IoT and Deep Learning," *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2020,pp. 708-713, doi: 10.1109

[13] M. Birket- Smith, N. Hasle and H. H. Jensen, "Electro dermal activity in anxiety disorders", *Acta Psychiatrica Scandinavica*, vol. 88, no. 05, pp. 350-355, 1993.

[14] J. Taelman, S. Vandeput, A. Spaepen and S. Van Huffel, "Influence of Mental Stress on Heart Rate and Heart Rate Variability", *4th European Conference of the International Federation for Medical and Biological Engineering (IFMBE)*

[15] A. Rana, "Stress among students: An emerging issue", *Rana | Integrated Journal of Social Sciences*, Jul. 2019

[16] Ahuja, Ravinder, and Alisha Banga. "Mental stress detection in university students using machine learning algorithms." *Procedia Computer Science* 152 (2019): 349-353.

[17] Pise, Anil Audumbar, et al. "Methods for facial expression recognition with applications in challenging situations." *Computational intelligence and neuroscience* 2022.1 (2022): 9261438.