FACIAL ACTION UNIT DETECTION

Author: Rahul. K, *IIITDM Kancheepuram, Chennai.*

ABSTRACT: This paper proposes a novel approach to detecting Facial Action Units (FAUs) using a Vision Transformer model, aiming to enhance the understanding of emotional expression and micro-emotions in psychological research. Utilizing the DISFA dataset, which includes videos with annotated facial action units, the study preprocesses the data by reducing frame rates, converting to grayscale, and applying contrast enhancement techniques. A pre-trained Vision Transformer, fine-tuned with additional dense layers and dropout, serves as the core model for multi-label classification of FAUs. The model is trained using the Adam optimizer and focal loss to address class imbalance, achieving strong performance metrics, including high precision-recall AUC scores.

The research highlights the challenges of handling class imbalance, limited dataset availability, and computational demands, and discusses the model's potential applications in stress analysis and micro-emotion detection. Future work will focus on validating the model on additional datasets and refining its capabilities for real-world applications.

KEYWORDS: Facial Action Units (FAUs), Vision Transformer, Micro-emotion detection, Multi-label classification, Psychological research, Emotion analysis, Stress detection, DISFA dataset, Image preprocessing, Class imbalance, Focal loss, TensorFlow, Computer vision, Pre-trained models, Action unit detection.

1 INTRODUCTION

Facial Action Unit (FAU) detection is a critical task in the fields of computer vision and psychology, as it enables the analysis of facial expressions to infer underlying emotional states and detect subtle micro-expressions. Understanding and interpreting facial expressions is fundamental for a range of applications, from psychological research and human-computer interaction to healthcare and security. In psychological research, FAU detection provides insights into human emotions, allowing for the study of affective states and social signals. For human-computer interaction, accurately detecting and interpreting facial expressions can enhance the ability of systems to respond to human emotions, leading to more intuitive and empathetic interfaces. In healthcare, FAU detection can assist in diagnosing mental health conditions, monitoring patient well-being, and improving the quality of life for individuals with emotional and behavioral disorders.

Traditional methods for FAU detection often rely on handcrafted features and classical machine learning algorithms. These methods typically involve manually designed descriptors such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Gabor filters, which capture specific facial features like edges, textures, and shapes. While these handcrafted features have been useful in identifying basic facial movements, they may lack the capability to fully capture the complexity and variability of facial expressions. Moreover, these traditional approaches require significant domain expertise for feature engineering and may not generalize well across different individuals, lighting conditions, and facial expressions.

In recent years, deep learning techniques, particularly convolutional neural networks (CNNs), have achieved significant success in image-based tasks due to their ability to automatically learn relevant features from large datasets. CNNs have been applied to FAU detection with promising results, as they can learn hierarchical feature representations from raw images, reducing the need for manual feature extraction. However, CNNs have limitations in effectively capturing long-range dependencies and contextual information within images. Facial expressions are composed of intricate and often subtle muscle movements that require a model to understand both local and global context to accurately detect and classify action units. This is particularly important for detecting micro-expressions, which are brief and involuntary facial expressions that reveal genuine emotions.

To address these challenges, this paper proposes the use of a Vision Transformer (ViT) [1] model for FAU detection. The Vision Transformer, a deep learning architecture based on the transformer model, leverages self-attention mechanisms to capture global contextual information and relationships between different parts of an image. Unlike CNNs, which rely on convolutions to extract local features, the ViT model processes images as sequences of patches and uses self-attention to model the interactions between all patches, allowing it to capture long-range dependencies and global patterns effectively. This ability to understand the overall structure of facial expressions, as well as the subtle nuances of individual action units, makes the Vision Transformer well-suited for the intricate task of identifying multiple facial action units simultaneously.

The proposed approach is evaluated on the DISFA (Denver Intensity of Spontaneous Facial Action) dataset, a widely used benchmark for FAU detection. The DISFA dataset provides a comprehensive collection of video sequences depicting spontaneous facial expressions, annotated with the intensity of 12 facial action units. This dataset presents a challenging testbed due to its diverse range of expressions, varying intensity levels, and inherent class imbalance, which reflects real-world scenarios where some action units are more frequent than others. To enhance data quality and mitigate class imbalance, we preprocess the dataset using techniques such as data augmentation, resampling, and balancing strategies that ensure each action unit is adequately represented during training.

Our Vision Transformer model, pre-trained on ImageNet and fine-tuned with additional dense layers, demonstrates superior performance in detecting action units across diverse facial expressions. By fine-tuning the model on the DISFA dataset, we adapt the pre-trained Vision Transformer to the specific task of FAU detection, optimizing it to recognize the subtle muscle movements associated with each action unit. The findings of this study highlight the potential of Vision Transformers in advancing the field of emotion recognition and micro-emotion analysis. By effectively capturing both local and global features, the ViT model offers a powerful tool for understanding complex facial expressions and detecting stress-related cues.

Furthermore, this research explores the application of FAU detection for stress detection, an area of growing interest in both academic and clinical settings. Stress is a common physiological and psychological response to various stimuli and can have significant health implications if not properly managed. Detecting stress through facial expressions offers a non-invasive method for monitoring emotional well-being and providing timely interventions. By leveraging the capabilities of Vision Transformers, this study aims to contribute to the development of robust models for automatic stress detection, which could have far-reaching impacts in areas such as mental health monitoring, workplace safety, and personalized healthcare.

In summary, the integration of Vision Transformers into FAU detection represents a promising advancement in the field of emotion recognition. By capturing the intricate details and global context of facial expressions, Vision Transformers offer a novel approach to understanding human emotions and detecting stress, paving the way for more effective and empathetic human-computer interactions and healthcare solutions.

1.1 Background and Motivation

This subsection could provide an overview of Facial Action Unit (FAU) detection and its significance in various fields, such as psychology, human-computer interaction, and healthcare. It would also discuss the importance of accurate emotion recognition and stress detection.
Stress is a psychological and physiological response to perceived challenges or threats. Chronic stress has been linked to various health issues, including cardiovascular diseases, anxiety disorders, depression, and weakened immune function. Accurate detection and monitoring of stress levels are crucial for developing preventive strategies and therapeutic interventions.
Traditional methods of stress detection, such as physiological measurements (heart rate, cortisol levels) or self-reported questionnaires, can be invasive, subjective, or impractical in real-time situations. Using facial expressions to detect stress offers a non-invasive alternative, allowing for continuous monitoring without disrupting a person's natural environment or activities.
Real-time stress detection using FACS can be applied in various fields, such as workplace wellness programs, educational settings, and driver monitoring systems. It can help in identifying stress early, allowing for timely interventions to mitigate its adverse effects.
With the advancement of computer vision and machine learning technologies, automatic detection of AUs from facial images has become feasible. This enables scalable and accurate stress detection systems that can be integrated into various digital platforms. [5].

Facial expressions are one of the primary non-verbal cues through which emotions and psychological states like stress are communicated. According to Paul Ekman's research, certain facial expressions are universally recognized as indicators of specific emotions. For stress detection, the focus is often on micro-expressions or subtle, involuntary facial movements that reveal genuine emotional states, even when someone attempts to hide them.

Action Units Related to Stress - Specific AUs have been associated with stress, anxiety, and related emotional states. For example:

- **AU4 (Brow Lowerer):** Often linked to stress, concentration, or anger.
- **AU9 (Nose Wrinkler):** Can be associated with disgust or frustration, which might appear under stress.
- **AU10 (Upper Lip Raiser):** Sometimes associated with a negative emotional response, which can correlate with stress.
- **AU15 (Lip Corner Depressor):** Associated with sadness or discomfort, potentially indicating stress.
- **AU17 (Chin Raiser):** Often appears in conjunction with other AUs in expressions of sadness or tension.

Understanding these AUs and their combinations helps in building models for stress detection. By training algorithms to recognize these patterns, it's possible to create systems that can detect stress with high accuracy.

Integrating facial expression analysis with other data sources, such as physiological signals (heart rate,speech) or contextual information (e.g., workplace environment, task difficulty), enhances the robustness of stress detection systems. Multi-modal approaches help mitigate the limitations of relying solely on facial cues, especially in scenarios where facial expressions might be masked or ambiguous. [13]

1.2 Objectives of the Study

The objectives of this research are centered around the development and evaluation of a robust model for Facial Action Unit (FAU) detection using Vision Transformers (ViTs). The primary goal is to harness the capabilities of ViTs to create a model that can effectively capture both local and global features within facial images, thereby improving the accuracy and reliability of FAU detection. To achieve this, the research aims to pre-train the Vision Transformer model on a large-scale dataset like ImageNet and then fine-tune it specifically for FAU detection on the DISFA dataset. This involves addressing the significant challenges associated with class imbalance, which is common in FAU datasets, by implementing advanced data preprocessing techniques such as augmentation, resampling, and balancing to ensure a diverse and representative training set. Furthermore, the study seeks to optimize the computational efficiency of the Vision Transformer model to make it suitable for real-world applications where resources may be limited. By achieving these objectives, the research aims to contribute to the field of emotion recognition by providing a powerful tool for detecting subtle facial expressions and stress-related cues, ultimately enhancing applications in areas such as mental health monitoring, human-computer interaction, and healthcare.

2 Literature Review

The paper introduces ResiDen, a novel network designed to improve Facial Action Unit (FAU) detection, by combining the strengths of ResNet and DenseNet architectures. It addresses the challenge of FAU detection in uncontrolled environments and explores the utility of information transfer from a Facial Expression Recognition (FER) network.

ResiDen's architecture features dense blocks with residual connections to mitigate vanishing gradients, and it integrates expression features extracted from a network trained on the RAF-DB dataset. The paper reviews traditional FAU detection methods, distinguishing between model-driven and data-driven approaches, and discusses the limitations of existing datasets.[10]

The paper by Perveen and Mohan presents a method for spontaneous facial expression recognition in natural settings by focusing on configural features from areas of the face with significant movement. These features are used to identify facial action units (FAUs), which are then combined using a coding system based on subjective interpretation to recognize expressions. The approach aims to reduce misclassification by minimizing overlap of FAUs across different expressions. The method is evaluated on several datasets, including CK+, JAFFE, SFEW, AFEW, MAHNOB laughter, and UVA-NEMO smile datasets, for tasks like expression recognition in controlled and uncontrolled environments, laughter localization, and distinguishing between posed and spontaneous smiles. The proposed method outperforms existing approaches by effectively handling issues like scaling and pose variations in video frames. The paper suggests future work on assessing expression intensity in videos from unconstrained environments. [12].

The research investigates public speaking anxiety (PSA) among graduates during thesis defenses, utilizing a novel neural network architecture for predicting levels of depression, anxiety, and stress based on facial expressions. The study combines the Facial Action Coding System (FACS) with the Depression Anxiety Stress Scale (DASS) to create a new database and achieve high accuracy in emotional state classification. Results indicate the architecture can effectively differentiate between healthy subjects and those affected by emotional disorders, with potential applications in virtual psychology, therapy, and health diagnostics. [6].

This paper presents methods for collecting and analyzing physiological data during real-world driving tasks to determine a driver's relative stress level. Electrocardiogram, electromyogram, skin conductance, and respiration were recorded continuously while drivers followed a set route through open roads in the greater Boston area. Data from 24 drives of at least 50-min duration were collected for analysis. The data were analyzed in two ways. Analysis I used features from 5-min intervals of data during the rest, highway, and city driving conditions to distinguish three levels of driver stress with an accuracy of over 97% across multiple drivers and driving days. Analysis II compared continuous features, calculated at 1-s intervals throughout the entire drive, with a metric of observable stressors created by independent coders from videotapes. The results show that for most drivers studied, skin conductivity and heart rate metrics are most closely correlated with driver stress level. These findings indicate that physiological signals can provide a metric of driver stress in future cars capable of physiological monitoring. Such a metric could be used to help manage noncritical in-vehicle information systems and could also provide a continuous measure of how different road and traffic conditions affect drivers. [8]

Facial Expression Recognition (FER) is presently the aspect of cognitive and affective computing with the most attention and popularity, aided by its vast application areas. Several studies have been conducted on FER, and many review works are also available. The existing FER review works only give an account of FER models capable of predicting the basic expressions. None of the works considers intensity estimation of an emotion; neither do they include studies that address data annotation inconsistencies and correlation among labels in their works. This work first introduces some identified FER application areas and provides a discussion on recognised FER challenges. We proceed to provide a comprehensive FER review in three different machine learning problem definitions: Single Label Learning (SLL)- which presents FER as a multiclass problem, Multilabel Learning (MLL)- that resolves the ambiguity nature of FER, and Label Distribution Learning- that recovers the distribution of emotion in FER data annotation. We also include studies on expression intensity estimation from the face. Furthermore, popularly employed FER models are thoroughly and carefully discussed in handcrafted, conventional machine learning and deep learning models. We finally itemise some recognise unresolved issues and also suggest future research areas in the field. [4]
Ground truth annotation of the occurrence and intensity of FACS Action Unit (AU) activation requires great amount of attention. The efforts towards achieving a common platform for AU evaluation have been addressed in the FG 2015 Facial Expression Recognition and Analysis challenge (FERA 2015). Participants are invited to estimate AU occurrence and intensity on a common benchmark dataset. Conventional approaches towards achieving automated methods are to train multiclass classifiers or to use regression models. In this paper, we propose a novel application of a deep convolutional neural network (CNN) to recognize AUs as part of FERA 2015 challenge. The 7 layer network is composed of 3 convolutional layers and a max-pooling layer. The final fully connected layers provide the classification output. For the selected tasks of the challenge, we have trained two different networks for the two different datasets, where one focuses on the AU occurrences and the other on both occurrences and intensities of the AUs. The occurrence and intensity of AU activation are estimated using specific neuron activations of the output layer. This way, we are able to create a single network architecture that could simultaneously be trained to produce binary and continuous classification output.[7]

Transformer is widely used in Natural Language Processing (NLP), in which numerous papers have been proposed. Recently, the transformer has been borrowed for many computer vision tasks. However, there are few papers to give a comprehensive survey on the vision-based transformer. To this end, we give an in-depth review of the vision-based transformer. We conclude 15 articles covering transformers on image object detection, multiple object tracking, action classification, and visual segmentation. Furthermore, we summarize 6 related datasets for corresponding tasks as well as their metrics. We also provide a comprehensive experimental comparison to validate the strength of transformer-based methods. We provide a brief introduction to the transformer and its applications on computer vision tasks, which can help beginners to recognize it. [2]

3 Proposed Work

In this section, we present the proposed approach for automatic spontaneous facial expression recognition in the wild. Figure 1 presents the block diagram of the proposed approach. The framework consists of four major steps: (i) Face and landmark detection, (ii)Facial Image preprocessing , (iii) facial action unit (FAUs) recognition.

3.1 Face and Landmark detection

We employed the Dlib library for facial landmark detection and face localization. The Dlib 68-point shape predictor was utilized to detect facial landmarks accurately. This predictor provides a set of 68 coordinates representing key points on the face, including the eyes, eyebrows, nose, mouth, and jawline. The precise localization of these landmarks is crucial for our subsequent feature extraction and classification tasks.

For face detection, we used the Dlib model `res10_300x300_ssd_iter_140000` [11], which is a pre-trained Single Shot Multibox Detector (SSD) model. This model operates on a resolution of 300x300 pixels and is designed to efficiently detect faces in images. It outputs bounding boxes that define the location of faces in the input images, allowing for effective cropping and further analysis. The combination of the Dlib 68-point shape predictor and the SSD face detector ensures robust and accurate face and landmark detection [3], forming a strong foundation for our facial action unit analysis pipeline.

3.2 Facial Image Preprocessing

To standardize the input for our model and optimize computational efficiency, all images were resized to a fixed resolution of 224x224 pixels with three color channels (RGB). This resizing ensures consistency in input dimensions, allowing the model to process images uniformly.

After resizing, we applied Contrast Limited Adaptive Histogram Equalization (CLAHE) [9] to each image. CLAHE is an advanced histogram equalization technique that enhances the contrast of images, particularly useful for improving

the visibility of features in facial regions. In our study, we experimented with various clip limits and grid sizes to optimize the contrast enhancement. The clip limit parameter controls the contrast enhancement degree by limiting the maximum slope of the cumulative distribution function, while the grid size determines the size of the contextual regions used for histogram equalization. By adjusting these parameters, we tailored the preprocessing step to maximize the visibility of facial landmarks and action units, ultimately improving the model's feature extraction and classification performance.

3.3 Facial Action Unit Recognition

For feature extraction, we utilized a pre-trained Vision Transformer (ViT) model, specifically the **Google/vit-base-patch16-224**. This model was fine-tuned by adjusting the dense layers to better suit our specific multi-label classification task. The Vision Transformer, trained on ImageNet weights, provided the best results, demonstrating its effectiveness in capturing intricate patterns and features in images.

To optimize the performance of the pre-trained Vision Transformer model (`google/vit-base-patch16-224`), we employed global average pooling (GAP) to reduce the number of parameters and features. GAP computes the average of each feature map, effectively reducing its dimensionality while retaining the global spatial information. This method helps to mitigate overfitting by limiting the model's complexity and ensuring that only the most relevant features are retained.

The Vision Transformer model leverages self-attention mechanisms to capture global context and relationships between different parts of an image. Unlike traditional convolutional neural networks, which focus on local receptive fields, the self-attention mechanism in ViTs allows the model to consider all regions of an image simultaneously, providing a holistic understanding of the visual input. This approach has proven highly effective for various image classification tasks, and it has been widely adopted in competitions and benchmarks, consistently achieving state-of-the-art results.

Following the global average pooling layer, we introduced multiple fully connected (dense) layers with dropout regularization. Dropout is applied to prevent overfitting by randomly setting a fraction of input units to zero during training, which forces the network to learn more robust features that generalize well to unseen data. The final fully connected layer consists of 12 neurons corresponding to the 12 output labels, representing the different action units detected in the facial images.

This combination of global average pooling and fully connected layers with dropout allows our model to effectively map the extracted features to the output labels while maintaining a balance between model complexity and generalization ability.

Furthermore, the architecture of the Vision Transformer inherently supports multi-label classification tasks. This capability is particularly advantageous for our study, as it enables the model to detect multiple action units simultaneously in a single image, aligning well with the nature of our dataset and the complexity of facial action unit detection.

Training Process:

- Trained on Kaggle GPU P100
- Framework used : Tensorflow
- Optimizer : Adam
- Batch size : 128
- Loss Function : Focal Loss (to handle class imbalance)
- Epochs : 80
- Train, validation, Test split : 52:13:35 percent.

4 Results of Training and Testing

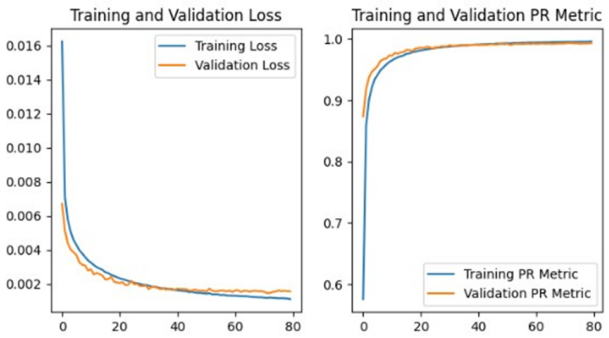

Figure 1: Training loss and pr-auc graphs.

AU	Precision	Recall	Accuracy	F1 Score	Muscle Movements
AU1	0.957679	0.925462	0.991087	0.941295	Inner Brow Raiser
AU12	0.973520	0.959785	0.982237	0.966604	Lip Corner Puller
AU15	0.945827	0.946174	0.992488	0.946000	Lip Corner Depressor
AU17	0.934609	0.933985	0.984950	0.934297	Chin Raiser
AU2	0.964017	0.951229	0.994497	0.957581	Outer Brow Raiser
AU20	0.942149	0.937381	0.995162	0.939759	Lip Stretcher
AU25	0.982657	0.974512	0.982557	0.978568	Lips Part
AU26	0.967082	0.949775	0.981639	0.958350	Jaw Drop
AU4	0.980817	0.953656	0.985399	0.967046	Brow Lowerer
AU5	0.929804	0.857749	0.995034	0.892325	Upper Lid Raiser
AU6	0.960855	0.967001	0.987319	0.963918	Cheek Raiser
AU9	0.979313	0.953666	0.995797	0.966320	Nose Wrinkler

Figure 2: metrics of each action unit

5 Iterative Experimentation and Findings

- Sampled 1 out of 4 frames to reduce dataset size (20fps to 5fps) but resulted in dataloss ,hence used all frames which have action unit in them to avoid action unit loss which are already less in number.Frames from the dataset got dropped which have action unit detected ,details of frame numbers dropped are given here click here.
- Pre trained Models like InceptionV3 , MobileNet and Efficient net are used to extract features from the facial images but their architecture supports single label classification better when compared to multi label classification and these models are outperformed by vision transformer.
- We have tried using facial landmarks and distance between facial landmarks and configural features to train but they didnt yeild good results.
- Traditional Feature extraction techniques such as LBP,HOG(histogram of oriented gradients) are used and they performed well by producing a smooth PR-AUC curve but vision transformer features performed better.
- Mediapipe Face extractor was used and couldnt identify faces accurately so we have used a better face extractor[11] and dlib's landmark predictor.
- Class imbalance was a serious issue while training and handling that with class weights was not easy so had to change the loss function to Focal Loss and also PR-AUC Metric is used to evaluate model's performance to get the trade off between precision and recall which indicate the level of False Negatives and False Positives which are important in stress analysis as False Negative is a case where we are predicting not stressed even if the person is stressed.

6 SUMMARY

This paper presents a novel approach to detecting Facial Action Units (FAUs) using a Vision Transformer (ViT) model, enhancing the understanding of emotional expressions and micro-emotions for applications in psychological research and human-computer interaction. Utilizing the DISFA dataset, we preprocess the facial images by resizing them to 224x224 pixels, converting to grayscale, and applying Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance image contrast. The face and landmark detection are performed using the Dlib library, with the 68-point shape predictor and a pre-trained Single Shot Multibox Detector (SSD) model for robust facial localization.

For feature extraction, we employ a pre-trained Vision Transformer (`google/vit-base-patch16-224`), fine-tuning it by adding dense layers optimized for multi-label classification of FAUs. The model leverages self-attention mechanisms to capture global context within images, demonstrating superior performance in detecting action units across diverse facial expressions. To prevent overfitting, global average pooling is used to reduce feature dimensionality, followed by fully connected layers with dropout regularization. The model is trained using the Adam optimizer and focal loss to address class imbalance, achieving high precision-recall AUC scores.

Our approach showcases the potential of Vision Transformers in advancing emotion recognition and micro-emotion analysis, highlighting the model's adaptability and effectiveness in handling multi-label classification tasks. Future work will focus on validating the model on additional datasets and refining its capabilities for real-world applications, such as stress detection and psychological assessment.

7 Challenges and Considerations

- Main challenge is handling class imbalance in multi label classification where we have so much frames which have no action unit.
- Dataset Availability , as there are no publicly available datasets for FACS and There are minimal datasets which are free for researchers.
- Dataset has very less occurrences of each action unit and even creating a personal dataset is not easy until one has expertise in Facial action coding system.
- Computational challenges : Training the model on entire dataset of high resolution images (792x1024) ,which was resolved after extracting face from image and
- resizing them ,yet was difficult to train on local machine for a model with such huge parameters.
- Interpreting the model's prediction as above video and csv files where we have binary predictions as well as Probabilities given.
- Plotting probabilities of action units against time with a threshold line to analyse the times when model predicted yes for an action unit

8 Conclusion

- Facial action unit detection has capability to capture micro emotions(appear at frame level – $1/25^{th}$ of a second).
- It is useful for stress analysis where the emotion of the person is captured at frame level.
- Accurate detection of facial action units can support research in psychology, helping to understand how emotions are expressed and perceived
- Will validate on other datasets to assess model's performance.

References

[1] Baroni, G., Rasotto, L., Roitero, K., Tulisso, A., Di Loreto, C., and Della Mea, V. (2024). Optimizing vision transformers for histopathology: Pretraining and normalization in breast cancer classification. *Journal of Imaging*, 10(5):108.

[2] Bi, J., Zhu, Z., and Meng, Q. (2021). Transformer in computer vision. In *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pages 178–188.

[3] Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685.

[4] Ekundayo, O. S. and Viriri, S. (2021). Facial expression recognition: A review of trends and techniques. *IEEE Access*, 9:136944–136973.

[5] Everton, M. D. (2020). Active system grounding with a novel distribution transformer design. *IEEE Open Access Journal of Power and Energy*, 7:183–190.

[6] Gavrilescu, M. and Vizireanu, N. (2019). Predicting depression, anxiety, and stress levels from videos using the facial action coding system. *Sensors*, 19.

[7] Gudi, A., Tasli, H. E., den Uyl, T. M., and Maroulis, A. (2015). Deep learning based facs action unit occurrence and intensity estimation. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–5.

[8] Healey, J. and Picard, R. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166.

[9] Hu, Y., Zeng, X., Huang, Z., and Dong, X. (2021). A preprocessing method of facial expression image under different illumination. In *2021 13th International Conference on Communication Software and Networks (ICCSN)*, pages 318–322.

[10] Jyoti, S. and Dhall, A. (2018). Expression empowered residen network for facial action unit detection. *CoRR*, abs/1806.04957.

[11] Matthias, D. and Managwu, C. (2021). Face mask detection paper.

[12] Perveen., N. and Mohan., C. K. (2020). Configural representation of facial action units for spontaneous facial expression recognition in the wild. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 93–102. INSTICC, SciTePress.

[13] Vetrekar, N., Ramachandra, R., Raja, K., and Gad, R. (2018). Detecting glass in ocular region based on grassmann manifold projection metric learning by exploring spectral imaging. In *2018 14th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, pages 106–113.