RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2614

Cross-Domain Transfer Learning with Vision Transformers for Fatigue Recognition from Facial Expressions

Rachana Yogesh Patil [a,*], Yogesh H. Patil [b], Apaprna Bannore [c], Deepali Nilesh Naik [d], Jotiram K. Deshmukh [e]

[a] Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India
[b] D Y Patil College of engineering, Akurdi, Pune, Maharashtra, India.
[c] SIES Graduate School of Technology, Nerul, Navi Mumbai, Maharashtra, India
[d] Symbiosis Centre for Management and Human Resource Development, Symbiosis International (Deemed University), Pune, India
[e] Bharati Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India
* Corresponding Author Email: rachana.patil@pccoepune.org

DOI: https://doi.org/10.54392/irjmt2614
Received: 08-06-2025; Revised: 16-11-2025; Accepted: 20-12-2025; Published: 13-01-2026

Abstract: Fatigue detection plays a critical role in mission-critical environments such as defense operations, transportation, and industrial control, where sustained attention and alertness are paramount for safety, operational efficiency, and human performance. This study introduces a real-time, non-intrusive fatigue detection system that employs Vision Transformers (ViTs) to identify subtle facial cues associated with fatigue. Unlike traditional methods that rely primarily on overt indicators such as yawning, eye closure, or head nodding, our approach leverages advanced deep learning techniques to capture nuanced micro-expressions and subtle behavioral patterns that are often overlooked. By applying transfer learning from the FER-2013 dataset to the NTHU-DDD dataset, we achieve enhanced model generalization, with ViT-L16 and R50+ViT-B16 architectures attaining classification accuracies of 70.31% and 72.79%, respectively. Visual attention maps reveal that FER-FFR models focus more effectively on critical facial regions, enabling precise feature extraction and interpretability. Furthermore, we introduce a fatigue level indicator that quantifies fatigue progression over consecutive video frames, demonstrating behavior that closely aligns with human fatigue dynamics. The proposed system is robust, scalable, and suitable for deployment in real-world operational settings, providing an automated, reliable, and objective solution for continuous fatigue monitoring, potentially enhancing safety, productivity, and decision-making in high-stakes environments.

Keywords: Facial Expression Recognition (FER), Fatigue Detection, CNN, Resnet, Attention Based Models

1. Introduction

In any organization, where the quality of the output depends on human performance, fatigue can cause performance degradation. Despite work quality, fatigue also poses a health risk to individuals [1]. Sleep disorders, depression, and stress result from fatigue, which can escalate and prove to be fatal [2]. Therefore, monitoring and managing fatigue is essential for safeguarding the health and safety of individuals in the workplace.

Aviation and particularly military aviation, is a challenging and demanding job that requires utmost professionalism. Slight mistakes cause an invaluable loss of trained human resources and costly warfighting equipment to the nation [3]. Fatigue monitoring through self-declaration has been in place for aircrew to avoid flying accidents/incidents and meet statutory requirements. However, it is often neglected for the ground crew involved in operations from the ops room or in combat enabling ground operation. Increased fatigue levels for them can prove to be detrimental towards achieving mission objectives. Fatigue may also affect the aviation unit's readiness, such as command and control centers, which are manned 24X7 and are mandated to respond to any aerial threats within seconds.

The present system of fatigue detection in aviation is manual. Officers responsible are supposed to ensure that air-warriors placed below him are well rested and not fatigued during the job [4]. However, the manual system has many flaws, such as worker to supervisor ratio, the strenuous task of fatigue detection, the set- ting of fatigue in monitoring officer himself, etc. It has failed on numerous occasions in the past. During hostility or heightened national security situation, the problems in

fatigue monitoring increase exponentially. Hence there is a need for automatic detection of fatigue.

Fatigue can be detected automatically using different sensors like electromyography (EMG), electrocardiogram (ECG), electroencephalograph (EEG), and camera [5]. The camera is a highly promising sensor for fatigue detection, as it provides multiple informative visual cues while being non-intrusive and not requiring the subject to wear any device. The human face is a powerful channel of non-verbal communication. It conveys a lot of semantic information such as age, gender, ethnicity, and emotion. Every human emotion, including fatigue, is instantly visible on the face. Also, it is challenging to disguise or mislead one's own face. The human face serves as the ideal candidate for the automatic detection of Automatic detection of fatigue from the face has been limited under Driver Drowsiness Detection (DDD) terminology [6]. In our approach, we are trying to extend it to a more general category of day-to-day settings tracking fatigue of people involved in the critical tasks. Inspired from the topic of FER, which has widely been researched in computer vision, we have used the terminology of Facial Fatigue Recognition (FFR).

Traditionally fatigue has been associated with closed eyes, yawning, and head movement [7]. Most of the FFR approaches have focused on the detection of the above signs on the face. The metric used included Percentage Eye Closure (PERCLOS) and Frequency of Mouth (FOM). However, these systems suffer from two main drawbacks: first is threshold needs to be set on these matrices, limiting the system's effectiveness and second is many other minute facial movements such as inner brow rise, outer brow rise, lip stretch and jaw drop have also been known to be markers for drowsiness.

Traditionally most of the deep learning-based approaches used CNNs for fatigue detection [8]. Humans perceive facial emotion or fatigue by looking at various important facial regions like eyes/lips/forehead, and ignore other details of hair, ear, etc. For detection, the model must attend to various regions of facial parts and Attention models provide a solution to this. We have used the Attention Based model in our approach.

It is also necessary for fatigue detection systems to quantify the level of fatigue. Mere classification into fatigue and non-fatigue may not be sufficient in real life scenario. We have taken motivation from bio-mathematical models for human performance under fatigue and provided rough estimate of fatigue level along with the detection.

As a part of this project, our key contributions can be summarized as follows. First, we propose a fatigue detection model that leverages a deep learning-based FER framework, which has been fine-tuned specifically for the FFR task. Second, our system not only detects the presence of fatigue but also monitors and estimates the level of fatigue over time, providing a more detailed and actionable assessment. Third, we integrate an attention-based modeling approach, allowing the system to focus on critical facial regions and capture subtle cues indicative of fatigue. Collectively, these contributions aim to provide a robust, accurate, and real-time solution for fatigue monitoring in critical operational settings.

1.1 Our Contributions

Our work is distinguished from existing FER and FFR approaches by the following key contribution

- Vision Transformer for FFR – Leveraging ViT models to capture subtle fatigue cues beyond the capability of CNNs.
- FER → FFR Transfer Learning – Reusing FER-learned features to improve FFR performance under limited data.
- Hybrid ViT Architecture – Combining ResNet-50 with ViT for enhanced feature extraction.
- Fatigue Level Quantification – Introducing a temporal indicator to monitor fatigue progression over time.

2. Related Work

The FER and FFR methods can be divided into two broad categories: conventional approaches in which features are handcrafted and deep learning-based approaches in which features are automatically generated through the output of a deep neural network.

2.1 FER Approaches

In conventional methods feature extraction step and the feature classification step are independent. Feature extraction is carried out using image processing techniques. Various methods for feature extractions include LBP (local binary pattern), gabor filter, scale-invariant feature transform (SIFT), speeded up robust features (SURF), and histogram of oriented gradients (HOG).

The features used in FER can be divided into three main types: geometric features, appearance features, and a combination of these two as hybrid features. ELLaban *et al* [9] implemented FER using SVM and KNN classifier. On their self-made dataset, SVM outperformed the KNN classifier. [10] Implemented FER using Fisher Independent Component Analysis (FLD) feature extraction and hidden markov model (HMM) for classification on CK+ dataset. Authors of [11] integrated geometric and textural features to form a hybrid feature vector during training. The hybrid feature improved the accuracy. Datta *et al*. [11] used an image segmentation-based approach using gabor filtering along with SVM

and showed that the image segmentation technique improved accuracy. Conventional approaches for FER are still being studied due to their relatively lower computing power and memory. However, most of the state-of-the-art FER techniques used deep learning based models.

2.2 Deep-Learning Based FER Approaches

Most DL-based FER approaches use CNNs as the backbone. FaceNet2ExpNet [12] proposed two-stage training from Face Recognition to FER. The fine-tuned face net serves as a suitable initialization for the expression net and is used to supervise the training of the convolutional layers.

Traditionally CNNs were directly employed on raw pixel intensities to learn the features. Some approaches tried to employ various hand-crafted features and their extensions as the network input rather than raw pixel data to strengthen the network's robustness to common distractions and to force the network to focus more on facial areas with expressive information [13].

Some approaches tried attention-based models, which help to suppress the contribution of surrounding deterrent elements and concentrates classification solely on facial regions. [14-16] proposed automatically learning the salient features for facial expressions using attention mechanisms. However, till the introduction of Vision Transformers, these attention-based models were used in conjunction with CNNs only.

Recently, Vision Transformers (ViTs) have gained attention in FER by modeling long-range dependencies across facial regions. Studies [17] demonstrate that ViTs outperform CNNs in capturing subtle intra-class variations, especially under occlusion and pose changes. These findings motivate our use of ViTs as a backbone for fatigue recognition.

2.3 FFR Approaches

Due to very little diversity in available FFR data (maximum number of subjects is less than 40), most of the FFR techniques have earlier used the conventional approach of hand-crafted features, and e.g., [18] used LBP + AdaBoost for FFR. Some approaches have used facial landmark detectors and relationships between those landmarks to create a feature vector for classification using SVMs. However, in FFR also, deep learning-based features proved to be more successful in real-world circumstances. Recently CNNs were used in [19] for fatigue detection. [20] used MTCNN pre-processing as a face detector and facial landmark detector towards fatigue detection and used regions of eye and mouth for FFR. Three-layered CNN has been used in [21] to capture various facial features for fatigue detection explicitly. The method did not focus only on eyes/mouth, rather used entire face regions to improve upon the previous approaches.

Most of the previous approaches were tested on the custom dataset, created specifically towards testing the approach used. On the NTHU-DDD dataset, [22] used three pre-trained deep neural networks (AlexNet, VGG-FaceNet, and FlowImageNet) along with two ensemble strategies (independently averaged architecture and feature-fused architecture) to classify each video frame as drowsy or not.

The approach showed that facial feature representation is essential for robust detection. However, the models like VGG-FaceNet, which are trained for face recognition, are dominated by face information due to the large-scale FR dataset, reducing the model's capability to learn facial expression discriminative features [23]. This reduces their effectiveness in FFR.

In 2004, the effect of facial expression on fatigue had been studied briefly in [24]. The approach designed the FFR system using the Facial Action Coding System (FACS), which is based on facial muscle changes and can characterize facial actions. Further to this, in 2010, [25], the FACS system was used to effectively discriminate between moderate and acute drowsiness. Both these approaches confirmed that whole face and facial features are essential for robust fatigue recognition. The domain closely related to FACS is facial expression recognition. Thus we have decided to model these critical facial features through transfer learning from the deep learning based FER model.

In the context of fatigue detection, multimodal approaches that integrate EEG and visual signals [26] and temporal deep learning architectures such as LSTMs and TCNs [27] have been proposed to capture temporal fatigue dynamics. However, transformer-based models for FFR remain underexplored, with only limited work attempting to leverage ViTs for fatigue state recognition. Moreover, existing studies do not address domain transfer between FER and FFR datasets, leaving a key research gap that our work aims to fill.

Table 1 provides a comparative summary of existing FER and FFR approaches. As seen, most FFR methods rely either on CNN-based pipelines or handcrafted features, while only limited works have begun exploring transformer architectures or multimodal inputs. Crucially, none of the prior studies have investigated transferring FER-learned representations into FFR tasks using Vision Transformers, nor combined hybrid CNN–ViT architectures with fatigue-level quantification. This gap forms the basis of our proposed contributions.

Table 1. Summary of Existing FER and FFR Approaches

Referene	Task	Dataset	Model/Method	Key Contribution	Limitations
[9], [10]	FER	CK+	Handcrafted features (SIFT, HOG, ICA) + SVM/HMM	Early handcrafted FER	Limited generalization
[12]	FER	FER2013	FaceNet2ExpNet (CNN)	Two-stage CNN transfer learning	Sensitive to occlusion
[14–16]	FER	FER2013, CK+	Attention-CNNs	Salient feature learning	Still CNN-constrained
[17]	FFR	Custom	LBP + AdaBoost	Early handcrafted FFR	Poor scalability
[19, 20]	FFR	NTHU-DDD	CNN + MTCNN	Eye & mouth region focus	Ignores global face cues
[21]	FFR	NTHU-DDD	AlexNet, VGG-FaceNet, FlowImageNet	Ensemble for fatigue	Limited to CNNs
[23, 24]	FFR	Custom	FACS-based	Muscle action coding	Manual coding intensive

3. Proposed Methodology

The proposed methodology for fatigue detection consists of a systematic pipeline that involves image preprocessing, feature extraction using a ViT model, and fatigue level quantification. The entire process can be divided into the following stages, as described in figure 1. The table 2 summarized all notations used in the proposed system.

3.1. Input Acquisition

The input to the system is a continuous video stream or a sequence of frames captured using a camera. Each frame is represented as a digital image, denoted as: $I \epsilon \mathbb{R}^{H \cdot W \cdot C}$. where H represents the height of the image, W represents its width, and C denotes the number of color channels. For most cases, RGB images are used, meaning C=3. The video stream is essentially a collection of these image frames captured over time, represented as $V = \{I_1, I_2, I_3 \ldots\ldots I_T\}$. The frame rate, typically measured in frames per second (FPS), determines how frequently the camera captures frames. Higher FPS values provide more granular motion data, which can be beneficial for fatigue detection systems. The acquired frames then serve as input for further processing, starting with face detection and subsequent fatigue analysis.

3.2. Face Detection and Cropping

MTCNN performs a facial analysis within each frame to detect both facial expressions and fatigue-related signs. MTCNN stands out as an optimal detection solution because it locates faces across different lighting conditions and rotated positions effectively. The algorithm detects faces together with alignment functions that deliver exact positions for facial landmarks.

After face detection the system removes the detected face from its original frame before resizing it to P×P pixels for processing. The processing step guarantees uniform frame dimensions for the ViT model input which optimizes its performance. The operation can be expressed mathematically through: $I_f = Resize(Crop(I))$. Through preprocessing the model can concentrate on facial features only while background elements become minimized which leads to enhanced computational efficiency.

3.3. Feature Extraction and Patch Projection (Hybrid ViT Approach)

The hybrid Vision Transformer (R50 + ViT-B16) model starts with a ResNet-50 backbone because it processes the input image before generating patches from raw images. The ResNet-50 operates as a strong feature extractor which generates a detailed intermediate representation from the input image. The model extracts the feature map from the 49th layer of ResNet-50 before its fully connected layer. This feature map captures hierarchical spatial and texture-based features useful for fatigue detection.

The extracted feature map is then divided into non-overlapping patches. During training, the ResNet-50 backbone is kept frozen to retain its strong low- and mid-level feature representations, while the ViT encoder layers and classification head are fine-tuned on the fatigue dataset. This selective fine-tuning ensures that the model adapts to fatigue-specific cues without overfitting to the small dataset size.

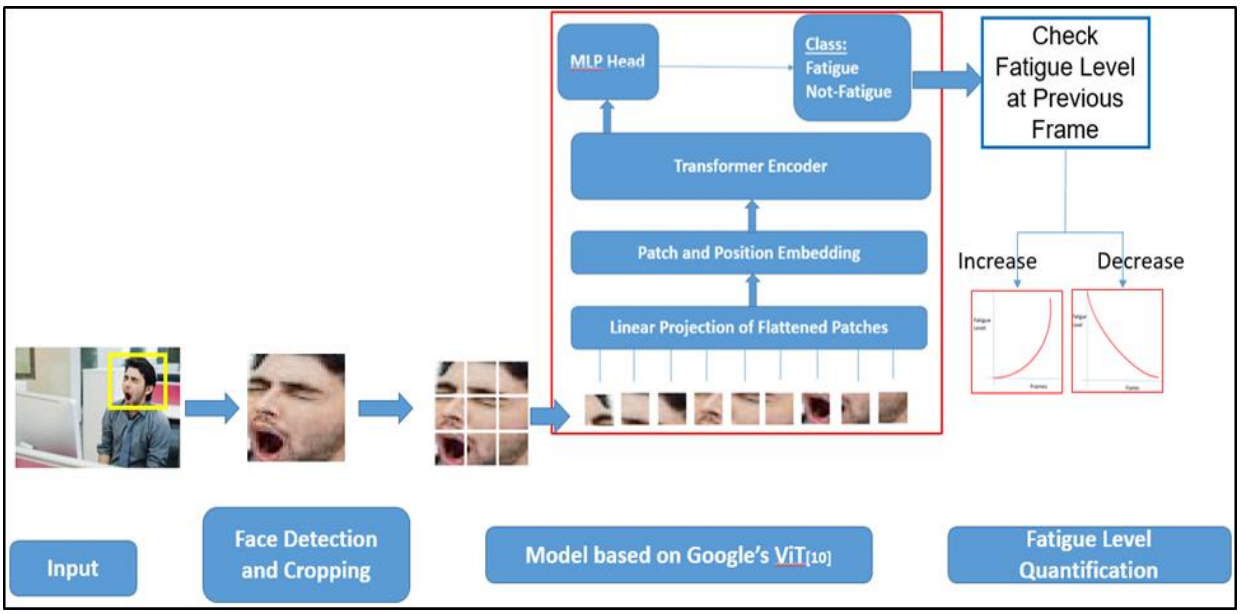

Figure 1. Architecture of Proposed system

Table 2. Notations used in proposed methodology

Notation	Meaning
I_t	Frame at time instance t
T	Total number of frames in the video
I_f	Cropped and resized facial image
I	Original input frame
$Crop(I)$	Operation to extract the facial region
$Resize$	Operation to resize the cropped image to P×P
z_0	Input sequence to the transformer
E_i	Positional embedding for the ith patch
N	Total number of patches
Q, K, V	Query, Key, and Value matrices derived from the input embeddings
d_k	Dimension of the key vectors
$\hat{y}$	Predicted class label
Z_L	Output from the final transformer layer
W_h, b_h	Learnable weights and biases of the MLP
σ	Activation function, typically softmax in the case of classification
F_t	Fatigue level at time t
α,β	Growth parameters
γ	Decay parameter

Each patch is flattened into a 1D vector, preserving spatial feature relationships. These vectors are passed through a trainable linear projection layer to map them into a fixed-size embedding space suitable for transformer input.

This hybrid approach allows the model to leverage, Convolutional inductive bias from ResNet-50, and global attention capabilities from the transformer. This step effectively transforms the feature-enriched spatial map into a sequence of patch embeddings ready for positional encoding.

3.4. Patch and Position Embedding

The models that we have used are mentioned as ViT-B 16, where the last number indicates the patch size. The feature map obtained from ResNet-50 is divided into fixed-size non-overlapping patches of size 16×16. Each patch is then flattened and linearly projected into a low-dimensional embedding space. To preserve the spatial arrangement of patches, a positional embedding is added to each patch embedding. The final input to the transformer is represented as: $z_0 = [x_1 + E_1, x_2 + E_2 \dots \dots \dots \dots x_N + E_N]$. For standalone ViT models, the transformer encoder is initialized from ImageNet-pretrained checkpoints. Fine-tuning is primarily applied to the higher transformer

layers and the classification head, as these layers capture task-specific variations such as subtle fatigue cues, while the lower layers remain frozen to preserve generic visual representations.

3.5. Transformer Encoder

A Transformer Encoder contains multiple stacked layers which process the embedded patches that form the input sequence. The encoder contains two main components per layer.

Multi-Head Self-Attention (MHSA) allows the model to understand relationships between all input patches by performing simultaneous attention across different sections of the information. The model can detect delicate facial indications that exist between different face regions through this mechanism. The model utilizes Feedforward Neural Network (FFNN) to perform independent non-linear transformations on each of its embeddings in order to boost its ability to represent information.

Our implementation relies on the ViT-Base (ViT-B16) architecture that contains 12 transformer encoder layers stacked on top of each other. Self-attention enables each transformer encoder layer to process embedded patches sequentially as it detects spatial dependencies between facial regions that matter for fatigue detection. The self-attention mechanism is mathematically expressed as: $Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$. Attention helps the model detect faint facial marks and expressions throughout different image areas so it can recognize indicators of fatigue.

The encoder system applies self-attention from NLP principles to vision tasks to attend relevant facial features throughout the face. The pre-training of ViT-Base on ImageNet21k and ImageNet2012 datasets allows efficient learning of fatigue recognition when applied to smaller training resources than traditional CNNs.

3. 6. MLP Head for Classification

The output from all transformer encoder layers reaches the [CLS] token for analysis through an MLP Head which performs the classification process. The final prediction of fatigue state comes from a Multi-Layer Perceptron that evaluates the learned representations. A binary classification approach is adopted, with two possible outputs: Fatigue and Not-Fatigue. The classification is performed using the following transformation:$\hat{y} = \sigma(W_h \cdot Z_L + b_h)$

The softmax function transforms the output logits into class probabilities: $P(y = c \mid x) = \frac{e^{z_c}}{\sum_j e^{z_j}}$. This results in a probability distribution over the two classes, enabling confident classification decisions.

The model used here is ViT-B16, which includes a lightweight MLP head at the end. This head is fine-tuned on the FFR dataset. Initially, the model is trained on FER data, helping it to learn generalized facial features. These learned features are then transferred and refined during the training on FFR, making the classification more robust. The transfer learning follows a two-stage process: first pre-training on FER-2013 for robust facial expression features, and then fine-tuning on the NTHU-DDD dataset for fatigue detection. This approach mitigates domain shift by gradually adapting learned features from general emotion cues to fatigue-specific signals.

This two-step transfer learning approach improves the model's ability to differentiate between subtle expressions and fatigue indicators.

3.7. Fatigue Level Quantification

The system needs to advance its real-world usability by both identifying fatigue existence in frames and measuring fatigue progression across time. Through this approach the system offers improved tracking with intervention opportunities at the right time.

The absence of properly annotated fatigue level datasets prevents supervised learning from being an effective approach for fatigue quantification. The system incorporates bio-mathematical models which defense forces and medical researchers use for fatigue prediction. The models derive their foundations from the sleep–wake pattern with selected versions implementing wearable devices which detect fatigue initiation through tracking of activity levels and awake duration.

We developed an effective method that utilizes the series of classification outcomes to generate fatigue measurements. Humans recover their energy during sleep and their energy level decreases naturally throughout waking hours. A continued series of fatigue classifications throughout multiple frames should trigger immediate action because it shows worsening fatigue levels.

The system follows up frame-wise classification by monitoring fatigue development throughout time. Continuous understanding of user alertness and avoidance of false positives require this essential feature.

- If a frame is classified as Fatigue, the fatigue level is incremented using an exponential growth function:$F_t = F_{t-1} + \propto \cdot e^{\beta t}$
- If the frame is classified as Not-Fatigue, the fatigue level decays exponentially: $F_t = F_{t-1} \cdot e^{-\gamma t}$

To ensure user safety, a threshold is defined. When the cumulative fatigue level exceeds this

predefined threshold, an alert is generated: $F_t > F_{threshold} \rightarrow Alert\ Generation$

The system uses this method to differentiate temporary fatigue indicators from actual alarms by detecting persistent patterns of fatigue. For implementation, we empirically selected the parameters α, β, and γ to balance growth and decay of the fatigue level in a realistic manner. Specifically, we set α = 0.5, β = 0.01, and γ = 0.02. These values were tuned through preliminary experiments to ensure that the fatigue level increases gradually during sustained fatigued states and decays at a reasonable rate during non-fatigue intervals. Importantly, the chosen values prevent the fatigue level from saturating too early at the maximum threshold or decaying to zero prematurely.

4. Experimental Results

4.1. Face Detection Evaluation

Face detector being the first part of the system dramatically affects the performance of the system. The requirement from the face detector is that it should be able to detect faces in an unconstrained environment and be fast for real-time performance. The two most widely available face detectors are VJ [28] and MTCNN [29]. We have implemented both the face detectors and compared their performance as shown in figure 2. VJ face detectors could not detect non-frontal faces and fail in the presence of multiple faces. MTCNN, on the other hand, worked well in an unconstrained setting and could detect non-frontal faces too. In a real-time setting, though MTCNN was found to be slower than VJ, it could meet the requirement of real-time face detection and could detect faces in 0.16 sec. Since we wanted the system to be employed in unconstrained settings, we decided to use MTCNN as a face detector.

We compared two widely used face detectors Viola-Jones (VJ) and MTCNN on key performance parameters. The graph shown in figure 3, comparing accuracy vs. face angle for VJ and MTCNN reveals that VJ performs excellently in frontal face detection (98%), but its accuracy drops sharply as the face angle increases, reaching 20% at 90°.

MTCNN, although starting with lower accuracy at 0° (85%), maintains more consistent performance across different angles, with 65% accuracy at 30° and 50% at 60°. This demonstrates MTCNN's better capability in detecting non-frontal faces, making it more suitable for real-time applications in unconstrained environments compared to VJ.

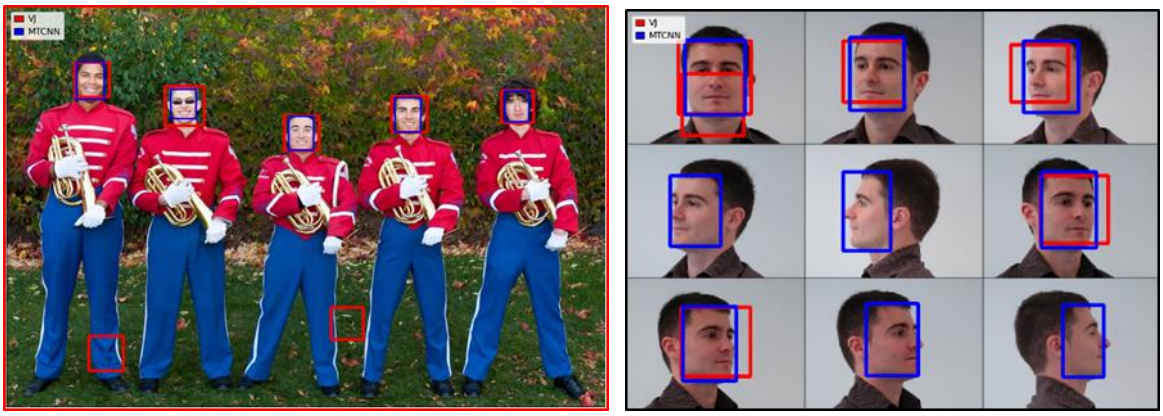

Figure 2. Performance of VJ and MTCNN

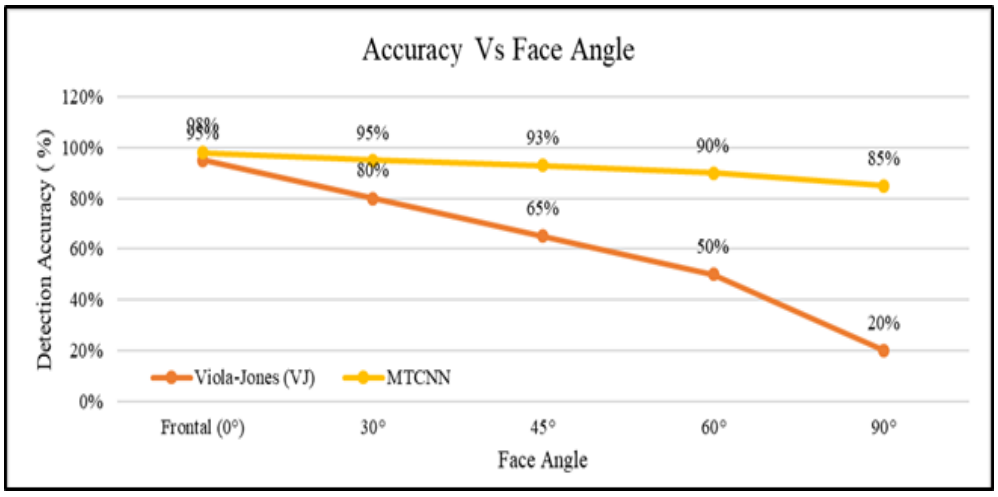

Figure 3. Comparison of Accuracy vs Face Angle for VJ and MTCNN

4.2 CNNs Vs Attention Based Models

Traditional deep-CNNs often struggle to capture fine-grained facial discriminative features due to subtle inter-class and intra-class variations, especially evident in the NTHU DDD dataset. Attention-based models, however, provide a more focused approach by attending to key facial regions that signal drowsiness or alertness. This capability is clearly demonstrated through occlusion maps. As shown in Figure 4, the Residual Attention Network (RAL) [30, 31] emphasizes meaningful facial regions associated with different drowsiness states, whereas the baseline CNN tends to focus on irrelevant, non-facial areas. Attention mechanisms not only improve model performance but also mimic the way humans interpret facial expressions related to fatigue and alertness.

The benefits of this focused learning become even clearer when we look at how these models learn over time. As shown in Figure 5, the RAL model not only learns faster but also achieves a lower training loss compared to the baseline CNN. This means it's picking up important features more efficiently and making fewer mistakes as training progresses.

But to really understand the performance difference, it helps to compare multiple models side by side. In Figure 6, we've plotted training and validation accuracy and loss for four different models: the baseline CNN, VGG16, ResNet18, and RAL. Here, RAL consistently comes out on top—it reaches the highest accuracy on the validation set and maintains the lowest validation loss. What's important is that it does so without overfitting, which is a common problem when models learn training data too well but fail on new examples. While VGG16 and ResNet18 show improvements over the baseline CNN, RAL offers a better balance between learning and generalization.

4.3 Performance of ViT Models

4.3.1 Training Details for Hybrid ViT Model

To ensure reproducibility and clarity, the implementation details of the hybrid R50 + ViT-B/16 model are summarized in table 3. The model was initialized with ImageNet-21k pretrained weights, and fine-tuning was carried out on the NTHU-DDD dataset. During training, the ResNet-50 backbone was frozen up to the 49th layer to preserve low-level convolutional features, while the transformer encoder and MLP classification head were fine-tuned for the fatigue detection task. Optimization was performed using the AdamW optimizer with cosine learning rate scheduling and early stopping to prevent overfitting. Data augmentation techniques such as random cropping, horizontal flipping, and brightness adjustment were applied to improve generalizability.

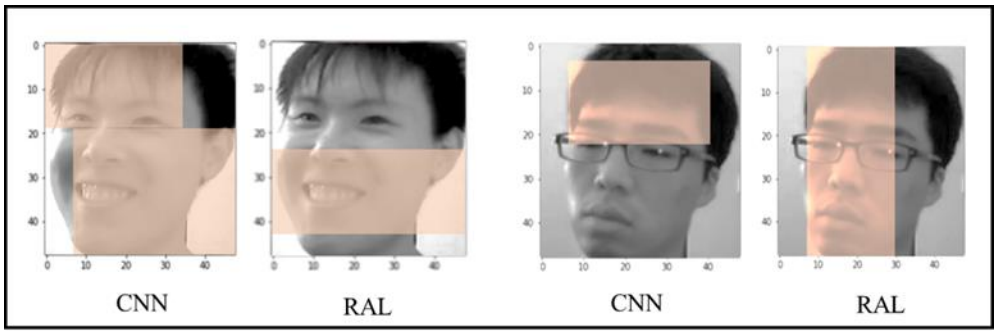

Figure 4. Occlusion maps highlighting attended regions by CNN and RAL

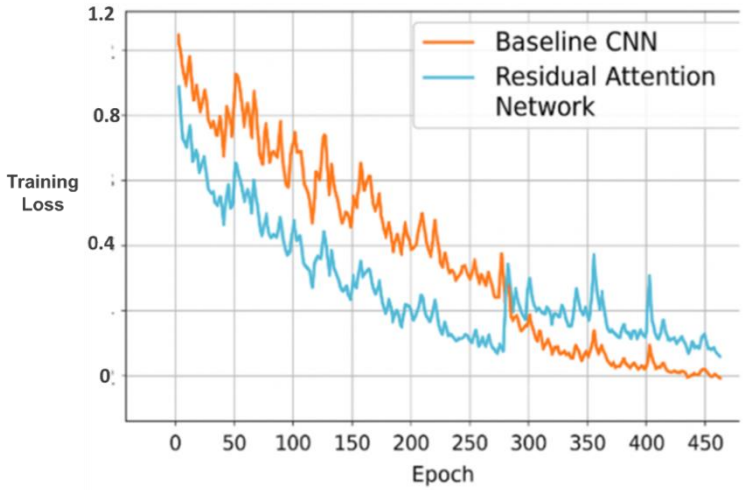

Figure 5. Training loss comparison between Baseline CNN and Residual Attention Network

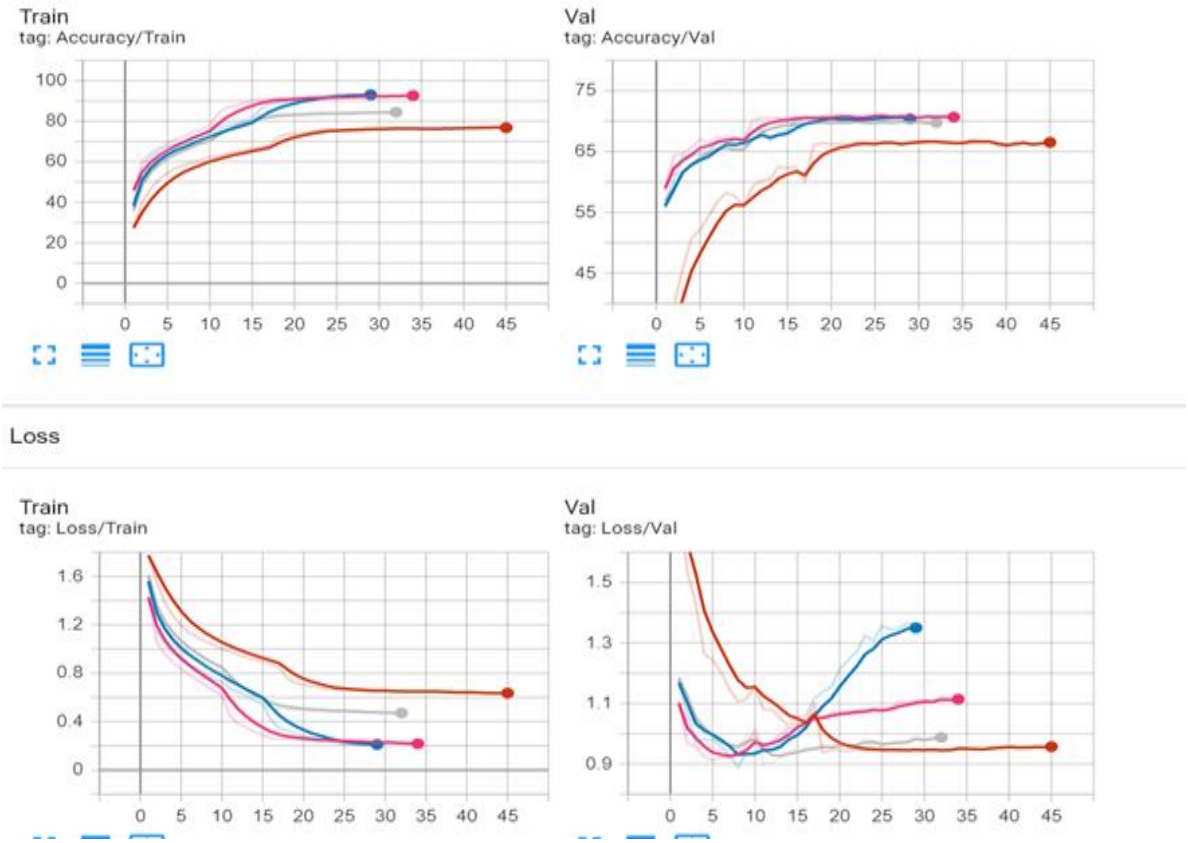

Figure 6. Training and validation accuracy and loss curves for multiple models

Table 3. Training Hyperparameters for Hybrid R50 + ViT-B/16 Model

Parameter	Value
Pretraining Weights	ImageNet-21k pretrained
Optimizer	AdamW
Initial Learning Rate	3×10^{-5} (cosine decay)
Batch Size	32
Epochs	50 (with early stopping)
Weight Decay	0.01
Data Augmentation	Random crop, horizontal flip, brightness adjustment
Fine-tuned Layers	Transformer encoder + MLP head

4.3.2 ViT Model for FER

We utilized pre-trained ViT models from Google's official checkpoints and modified them to suit the FER task. Various configurations were explored, including model size (Base, Large, Huge), patch size (16, 32), and input image dimensions.

Our experiments showed that Large (L) models consistently outperformed Base (B) models in terms of accuracy for the FER task. However, scaling further to Huge (H) models did not yield significant performance improvements, likely due to the need for extensive pre-training on large proprietary datasets, which was not feasible in our setup.

We also observed that a patch size of 16 resulted in better performance compared to a patch size of 32. To differentiate from the pre-training setup, we finalized the input image size as 224×224.

Interestingly, hybrid models like R50+ViT-B/16 demonstrated performance on par with the large ViT models. Based on these observations, we finalized the top two models — ViT-L/16 and R50+ViT-B/16 — as the source models for the subsequent FFR task.

The table 4 presents a comparative analysis of different ViT-based models evaluated on the NTHU DDD dataset in terms of Accuracy, Precision, Recall, and F1-score.

Among the models, ViT-L/16 achieved the highest test accuracy of 73.36%, followed closely by the hybrid model R50+ViT-B/16 with 72.33%, showing strong performance across all metrics. The base model ViT-B/16 also performed reasonably well with an accuracy of 71.60%, while ViT-H/14 showed similar results but without significant improvement despite increased model complexity. Notably, ViT-L/32, which uses a larger patch size, performed the worst among the ViT variants, indicating that finer patch granularity (patch size 16) captures better discriminative features for drowsiness detection. These results suggest that both model architecture and patch size play a critical role in the effectiveness of ViT models for driver monitoring applications.

4.3.3 ViT Model for FFR

The research evaluated the performance of ViT models for FFR by testing ViT-L 16 and R50+ViT-B 16 architecture in combination. The models received training and testing on NTHU-DDD database to determine their operational capacity when detecting driver drowsiness. All frames extracted from the NTHU-DDD dataset video recordings became part of the preprocessing phase. The dataset provided frame-by-frame fatigue state labels which researchers used to label each frame. Training involved eighteen subjects but testing exclusively employed the four remaining subjects to conduct independent subject testing.

As shown in table 5, the ViT-L 16 model delivered 70.31% accuracy during evaluation. The hybrid R50+ViT-B 16 model demonstrated higher performance than the other model by reaching 72.79% accuracy. The results demonstrate that using convolutional backbone networks such as the hybrid model leads to superior generalization capabilities for detecting subtle facial variations like fatigue.

4.3.4 FER-FFR Models

This section demonstrates how the facial feature understanding from a ViT-L 16-based FER model can be utilized for FFR tasks. The idea is to transfer discriminative facial cues from the FER-2013 dataset to the NTHU-DDD dataset. The FER-FFR model was trained using a combination of both datasets.

As a baseline, a standalone ViT-L 16 FFR model was also trained on the NTHU-DDD dataset. The standalone FFR model achieved an accuracy of 70.31%, while the FER-FFR model achieved 68.52%. To mitigate domain shift between the FER and FFR datasets, we fine-tuned the FER-pretrained model on NTHU-DDD and allowed the attention layers to re-focus on task-relevant fatigue cues. This adaptation ensures that general facial features learned from FER are effectively transferred and specialized for fatigue detection. Although there was a slight drop in accuracy, As summarized in table 6, the drop in accuracy is likely due to the small test set (only 4 subjects), and we expect the FER-FFR model to outperform in real-world, unconstrained settings.

Furthermore, the trade-off highlights an important distinction: while the standalone FFR model achieves marginally higher numerical accuracy, the FER-FFR model provides better generalization by focusing on meaningful facial regions, as observed in the attention maps. This indicates that transfer learning improves robustness across varying conditions, even if absolute accuracy on a small test set appears slightly lower.

5. Fatigue Level Indicator

To evaluate the performance of our fatigue level indicator, we first experimented by integrating it with randomized fatigue predictions. This initial setup allowed us to verify that the indicator did not saturate at the extremes (i.e., fatigue level 1 or 100). We observed the fatigue level dynamics through plotted graphs and fine-tuned the associated constant, which resulted in a reasonably responsive indicator that increased or decreased in accordance with the proposed fatigue quantification logic, as demonstrated in figure 7 (a).

Table 4. Performance Comparison of ViT-Based Models

Model Name	Accuracy	Precision	Recall	F1-Score
ViT-B/16	71.60%	70.80%	72.10%	71.44%
ViT-L/16	73.36%	73.90%	73.10%	73.50%
ViT-L/32	69.54%	68.40%	70.00%	69.21%
ViT-H/14	71.19%	70.50%	71.80%	71.14%
R50+ViT-B/16	72.33%	72.60%	72.00%	72.30%

Table 5. Performance of Selected ViT Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
ViT-L 16	70.31	69.85	70.10	69.97
R50+ViT-B 16	72.79	72.30	73.00	72.64

Table 6. Detailed performance comparison of ViT-L 16 on FFR and FER-FFR Tasks

Model Variant	Training Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Observations
ViT-L 16 (FFR)	NTHU-DDD only	70.11	71.20	69.80	70.49	Focused more on redundant regions (e.g., cheeks, background)
ViT-L 16 (FER-FFR)	FER-2013 + NTHU-DDD	69.52	69.40	67.80	68.59	Better coverage of facial features; generalizes well to real-world settings

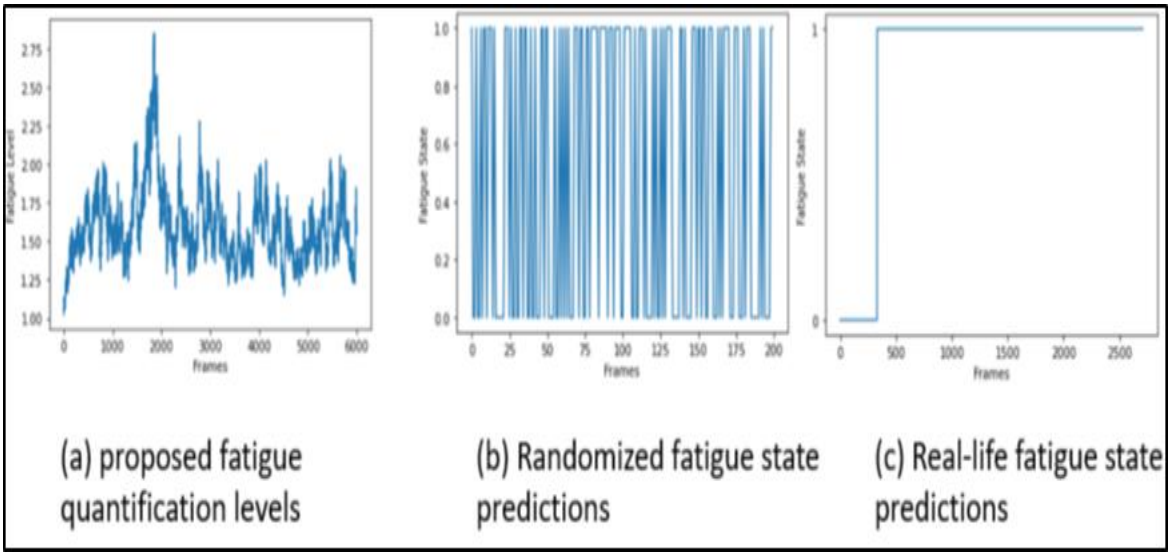

Figure 7. Fatigue Level Quantification

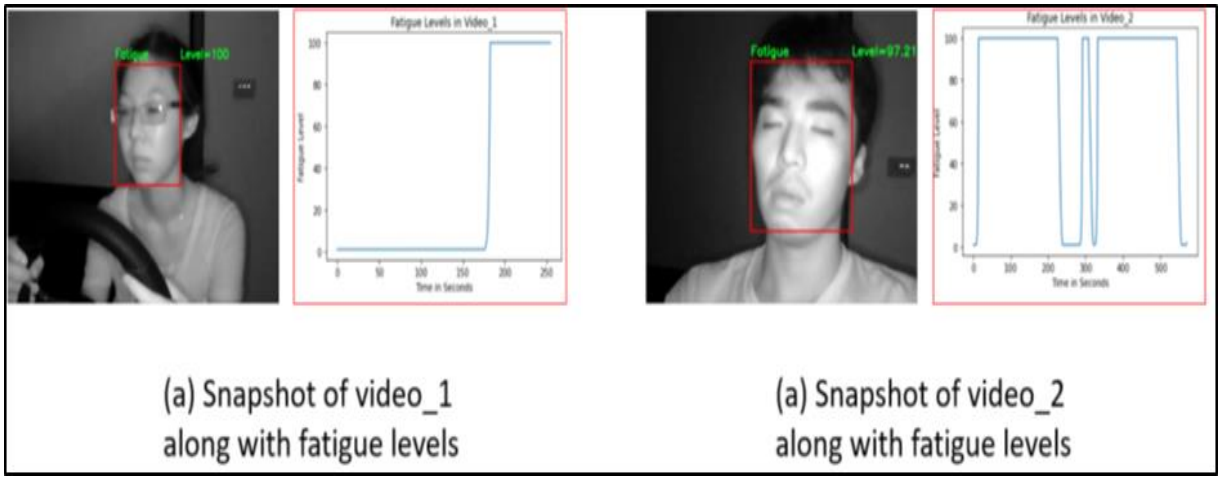

Figure 8. Snapshot of dataset videos along with fatigue level

However, a limitation of this approach was identified the fatigue level failed to reach the extreme value of 100. This behavior can be explained by considering real-life human behavior: the transition between fatigue and non-fatigue states is typically gradual and sustained. For example, once an individual enters a fatigued state, they generally remain fatigued for a period of time, and likewise for a non-fatigued state. The randomized prediction model lacked this temporal consistency and, consequently, was unable to generate long sequences of the same prediction. This shortcoming is clearly illustrated in figures 7(b) and 7(c).

To validate the indicator in a more realistic context, we selected two videos from our NTHU-DDD dataset, each providing frame-by-frame annotations of the subject's fatigue state:

- In the first video, the subject begins in a non-fatigue state and gradually transitions into a fatigue state.
- In the second video, the subject transitions from non-fatigue to fatigue and then back to non-fatigue.

The fatigue level plots for these videos showed satisfactory and intuitively accurate trends, as depicted in Figure 8(a) and 8(b). These plots provide strong evidence that our method effectively quantifies fatigue states and dynamically reflects the subject's level of alertness.

Although the current fatigue level estimation system is rudimentary, it offers a promising approach for transforming discrete fatigue states into a continuous fatigue level indicator. Such a metric could be highly beneficial in real-world deployments, enabling supervisors or safety systems to continuously monitor and assess the subject's fatigue level, and respond proactively to fatigue-related risks.

6. Conclusion

In this study, we proposed a real-time fatigue detection framework leveraging ViT models and transfer learning from facial expression recognition tasks. By using FER-trained ViT models for Fatigue Frame Recognition (FFR), we achieved competitive performance while also improving interpretability through attention maps. The use of the NTHU-DDD dataset validated the effectiveness of our approach, with ViT-L 16 and R50+ViT-B 16 showing promising accuracy. We also introduced a fatigue level quantification method that estimates fatigue severity over time using sequential predictions. This approach offers a robust, non-intrusive, and adaptable solution for monitoring fatigue in critical applications, addressing limitations of traditional fatigue detection methods and small datasets. While the proposed system demonstrates strong potential, it is not without limitations. The reliance on a relatively small test set restricts broader generalizability, and the fatigue level quantification remains a simplified approximation. Future work will explore larger and more diverse datasets, refine temporal modeling of fatigue progression, and incorporate multi-modal signals (e.g., physiological data) to enhance reliability in real-world deployments.

References

[1] C. Sutherland, A. Smallwood, T. Wootten, N. Redfern. Fatigue and its impact on performance and health. British Journal of Hospital Medicine, 84(2), (2023) 1-8. https://doi.org/10.12968/hmed.2022.0548

[2] D.N. Reinken, J. Rizek. Strategies to Prevent and Effectively Respond to Compassion Fatigue and Burnout. Journal of Emergency Nursing, 51(2), (2025) 205-210. https://doi.org/10.1016/j.jen.2024.10.004

[3] Y. Li, J. He. A review of strategies to detect fatigue and sleep problems in aviation: Insights from artificial intelligence. Archives of Computational Methods in Engineering, 31(8), (2024) 4655-4672. https://doi.org/10.1007/s11831-024-10123-5

[4] H. Pan, Y. Hu, Y. Wang, V. Duong. Fatigue Detection in Air Traffic Controllers: A Comprehensive Review. IEEE Access, 12, (2024) 185869 – 185880. https://doi.org/10.1109/ACCESS.2024.3513292

[5] P. Ertl, A. Kruse, M. Tilp. Detecting fatigue thresholds from electromyographic signals: A systematic review on approaches and methodologies. Journal of Electromyography and Kinesiology, 30, (2016) 216-230. https://doi.org/10.1016/j.jelekin.2016.08.002

[6] Y. Albadawi, M. Takruri, M. Awad. A review of recent developments in driver drowsiness detection systems. Sensors, 22(5), (2022) 2069. https://doi.org/10.3390/s22052069

[7] L. Dziuda, P. Baran, P. Zieliński, K. Murawski, M. Dziwosz, M. Krej, M. Piotrowski, R. Stablewski, A. Wojdas, W. Strus, H. Gasiul, M. Kosobudzki, A. Bortkiewicz. Evaluation of a fatigue detector using eye closure-associated indicators acquired from truck drivers in a simulator study. Sensors, 21(19), (2021) 6449. https://doi.org/10.3390/s21196449

[8] Z. Zhao, N. Zhou, L. Zhang, H. Yan, Y. Xu, Z. Zhang. Driver fatigue detection based on convolutional neural networks using em-CNN. Computational intelligence and neuroscience, 2020(1), (2020) 7251280. https://doi.org/10.1155/2020/7251280

[9] H.A. ELLaban, A.A. Ewees, A.E. Elsaeed. A real-time system for facial expression recognition using support vector machines and k-nearest

neighbor classifier. International Journal of Computer Applications, 159(8), (2017) 23-29.

[10] J.J. Lee, M.Z. Uddin, T.S. Kim. (2008) Spatiotemporal human facial expression recognition using fisher independent component analysis and Hidden Markov Model. In 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, Vancouver, BC, Canada, 2546-2549. https://doi.org/10.1109/IEMBS.2008.4649719

[11] S. Datta, D. Sen, R. Balasubramanian. (2017) Integrating geometric and textural features for facial emotion classification using SVM frameworks. In Proceedings of International Conference on Computer Vision and Image Processing: CVIP 2016, Springer Singapore, 1, 619-628. https://doi.org/10.1007/978-981-10-2104-6_55

[12] H. Ding, S.K. Zhou, R. Chellappa. (2017) Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), IEEE, Washington, DC, USA. https://doi.org/10.1109/FG.2017.23

[13] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, K. Yan. A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Transactions on Multimedia, IEEE, 18(12), (2016) 2528-2536. https://doi.org/10.1109/TMM.2016.2598092

[14] V. Mavani, S.Raman, K.P. Miyapuram. (2017) Facial expression recognition using visual saliency and deep learning. In Proceedings of the IEEE international conference on computer vision workshops,IEEE, Venice, Italy, 2783-2788. https://doi.org/10.1109/ICCVW.2017.327

[15] P.D. Marrero Fernandez, F.A. Guerrero Pena, T. Ren, A. Cunha. (2019) Feratt: Facial expression recognition with attention net. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, IEEE, Long Beach, CA, USA, https://doi.org/10.1109/CVPRW.2019.00112

[16] K. Wang, X. Peng, J. Yang, D. Meng, Y. Qiao. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, IEEE, 29, (2020) 4057-4069. https://doi.org/10.1109/TIP.2019.2956143

[17] A. Jarndal, H. Tawfik, A.I. Siam, I. Alsyouf, A. Cheaitou. A real-time vision transformers-based system for enhanced driver drowsiness detection and vehicle safety. IEEE Access, IEEE, 13, (2024) 1790-1803. https://doi.org/10.1109/ACCESS.2024.3522111

[18] Z. Liu, Y. Peng, W. Hu. Driver fatigue detection based on deeply-learned facial expression representation. Journal of Visual Communication and Image Representation, 71, (2020) 102723. https://doi.org/10.1016/j.jvcir.2019.102723

[19] B.K. Savaş, Y. Becerikli. 2020. Real time driver fatigue detection system based on multi-task ConNN. IEEE, Access, 8, (2020)12491-12498. https://doi.org/10.1109/ACCESS.2020.2963960

[20] Y. Ji, S. Wang, Y. Zhao, J. Wei, Y. Lu. Fatigue state detection based on multi-index fusion and state recognition network. IEEE Access, 7, (2019) 64136-64147. https://doi.org/10.1109/ACCESS.2019.2917382

[21] K. Dwivedi, K. Biswaranjan, A. Sethi. (2014) Drowsy driver detection using representation learning. In 2014 IEEE international advance computing conference (IACC), IEEE, Gurgaon, India, 995-999. https://doi.org/10.1109/IAdCC.2014.6779459

[22] S. Park, F. Pan, S. Kang, C.D. Yoo. (2016) Driver drowsiness detection system based on feature representation learning using various deep networks. In Asian conference on computer vision Cham: Springer International Publishing, 154-164. https://doi.org/10.1007/978-3-319-54526-4_12

[23] H. Ding, S.K. Zhou, R. Chellappa. (2017) Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), IEEE, Washington, DC, USA. https://doi.org/10.1109/FG.2017.23

[24] H. Gu, Q. Ji. (2004) An automated face reader for fatigue detection. In Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings. IEEE, 111-116.

[25] E. Vural, M. Bartlett, G. Littlewort, M. Cetin, A. Ercil, J. Movellan. (2010) Discrimination of moderate and acute drowsiness based on spontaneous facial expressions. In 2010 20th International Conference on Pattern Recognition IEEE, Istanbul, Turkey, 3874-3877. https://doi.org/10.1109/ICPR.2010.943

[26] S. Cao, P. Feng, W. Kang, Z. Chen, B. Wang. Optimized driver fatigue detection method using multimodal neural networks. Scientific Reports, 15(1), (2025) 12240. https://doi.org/10.1038/s41598-025-86709-1

[27] J. Bai, W. Zhu, S. Liu, C. Ye, P. Zheng, X. Wang. A Temporal Convolutional Network–Bidirectional Long Short-Term Memory (TCN-BiLSTM) Prediction Model for Temporal Faults in Industrial Equipment. Applied Sciences, 15(4), (2025) 1702. https://doi.org/10.3390/app15041702

[28] T.K. Soon, N.K. Ibrahim, B. Hussin, A. Idris, N.M. Yaacob, M.A. Algaet, N.A. Jalil. (2018) The utilization of feature based Viola-Jones method for face detection in invariant rotation. International Journal of Advanced Computer

Science and Applications, 9(12). https://dx.doi.org/10.14569/IJACSA.2018.091211

[29] R. Jin, H. Li, J. Pan, W. Ma, J. Lin. (2021) Face recognition based on MTCNN and Facenet. In AAAI Conference on Artificial Intelligence.

[30] R.Y. Patil, Y.H. Patil, S.U. Bhandari. FER to FFR: a deep-learning-based approach for robust fatigue detection. International Journal of Computer Applications in Technology, 72(3), (2023) 203-211. https://dx.doi.org/10.1504/IJCAT.2023.133292

[31] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang. (2017) Residual attention network for image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, Honolulu, HI, USA. https://doi.org/10.1109/CVPR.2017.683

Authors Contribution Statement

Rachana Yogesh Patil: Conceptualization, methodology, Software, validation, Formal analysis, Investigation, Writing - Original Draft. Yogesh H. Patil: Formal analysis, Investigation, Writing - Original Draft, Apaprna Bannore: Data Curation, Writing - Original Draft. Deepali Nilesh Naik: Writing - Review & Editing. Jotiram K. Deshmukh: Review & Editing. All the authors read and approved the final version of the manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes

About the License