Kufa Journal of Engineering
Vol. 17, No. 2, April 2026, P.P. 647-668
Article history: Received 17 May 2025, last revised 18 August 2025, accepted 23 August 2025

A DEEP LEARNING FRAMEWORK FOR REAL-TIME STRESS DETECTION USING FACIAL EXPRESSIONS

N.Nithya[1], A.Althaf Ali[2], S.Parvathi[3], and P.Mohamed Sajid[4]

[1] Department of Data Science, SRM Institute of Science and Technology, Ramapuram Campus, Chennai, India, Email:nithyamail2020@gmail.com.

[2] Department of MCA, Madanapalle Institute of Technology & Science (MITS), Andhra Pradesh, India, Email:althafalia@mits.ac.in.

[3] Department of CSE, Erode Sengunthar Engineering College, Perundurai , India, Email:5680sparvathi@gmail.com.

[4] Department of EEE, C. Abdul Hakeem College of Engineering and Technology, Melvisharam, India, Email:mohammed_sajid.ece@cahcet.edu.in.

https://doi.org/10.30572/2018/KJE/170239

ABSTRACT

Given its direct effects on human health productivity at work and general well-being stress detection has emerged as a crucial area of affective computing research. Self-reports and physiological sensors which are either subjective invasive or expensive to implement at scale are frequently used in traditional stress assessment methods. This study suggests an Enhanced VGG-Based Approach for Facial Expression-Driven Stress Detection (EVSD-Net) which uses subtle facial cues to accurately identify stress levels in order to overcome these limitations. By adding lightweight convolutional blocks to eliminate unnecessary parameters a hybrid channel–spatial attention mechanism to highlight stress-relevant areas (like wrinkles on the forehead tight lips and strained eyes) and a fine-tuned transfer learning strategy initialized with pretrained ImageNet weights the suggested framework enhances the standard VGG16 architecture. Both binary (stressed and non-stressed) and Softmax classifiers are supported by a hybrid classification layer that integrates fully connected (FC) layers. non-stressed) and multi-class stress classification (high medium and low). Preprocessing and augmentation of facial datasets improved feature extraction using the modified VGG network dimensionality reduction and regularization for better generalization and Softmax classification comprise the four main stages of the methodology. Accuracy Precision Recall F1-score and AUC metrics were used to evaluate the experimental results using four benchmark facial expression and stress-related

datasets. The results show that EVSD-Net outperforms baseline models like standard VGG16 ResNet18 and MobileNetV3 with an overall accuracy of 97.84%. The superiority of the suggested model is further confirmed by a comparison with previous research which shows significant improvements in detection accuracy and robustness under various illumination pose and demographic circumstances.

KEYWORDS

Stress detection, facial expression recognition, VGG, attention, multi-task learning, computer vision, affective computing.

1. INTRODUCTION

One of the most widespread issues affecting people's health and productivity is stress which is an unavoidable part of contemporary life. Chronic and uncontrolled stress has been repeatedly associated with a range of detrimental outcomes whereas acute stress can serve adaptive functions by improving alertness and enabling quick responses to immediate threats. These include long-term effects like cardiovascular disease depression and immune dysfunction as well as diminished cognitive function poor decision-making and burnout. Stress-related illnesses have already been recognized by the World Health Organization as significant contributors to the burden of disease worldwide. Persistent stress not only deteriorates individual well-being in occupational settings like healthcare education military service and customer service but it also compromises organizational efficiency and safety. Thus, there is a critical need for early accurate and scalable stress detection in both research and society. Traditionally, physiological signals like electrodermal activity, electrocardiography and related metrics such as heart rate variability, cortisol assays in bio fluids like blood or saliva, and electroencephalography were used for measurement of stress reactivity. There are significant downsides to these modalities for assessment in real-world context. First, many require physical sensors that either have to be attached directly to the body in different ways or involve the collection of samples from blood or saliva. These are often uncomfortable and not suitable for everyday use. Thirdly, privacy cost and compliance issues are frequently brought up by widespread deployment in public spaces like workplaces or schools. On the other hand, vision-based methods offer an appealing substitute for stress detection. Smartphones laptops cars and workplaces are already equipped with cameras which provide a passive inexpensive and non-intrusive way to monitor stress without making physical contact with the person. Particularly facial expressions are among the most potent markers of human emotional states. Happiness sadness anger fear disgust surprise and neutrality are just a few of the many emotions that the human face can express. These emotions frequently co-occur or overlap with stress.

Psychological research has demonstrated a correlation between increased stress or cognitive load and subtle muscular movements like lip pressing brow furrowing and eye narrowing. These cues in contrast to physiological signals can be continuously remotely and even retroactively recorded using video data. However, it is not easy to use facial expressions to identify stress. There are still two major issues that need to be addressed.

Facial expression recognition (FER) has advanced significantly as a result of deep learnings application to affective computing which has transformed computer vision. When it comes to identifying common emotional categories models built on convolutional neural networks

(CNNs) have outperformed conventional handcrafted feature methods. VGG-16 is still a popular and reliable backbone among CNN architectures. VGG architectures are still preferred in FER tasks because of their stable gradients simple design and efficient transfer learning capabilities from large-scale pretraining such as ImageNet or VGG-Face even though deeper networks like ResNet or more recent transformer-based models have shown higher accuracy in large-scale image classification. For learning hierarchical spatial features VGGs straightforward sequential stacking of convolutional layers offers a strong inductive bias. This is useful when fine-tuning on comparatively smaller FER or stress datasets. However, stress cues are complex and simple VGG networks are insufficient to capture this. Unfortunately, CNN-based approaches are unable to detect such fine-grained features that are more specific to identifying when a face may be showing stress. Additionally, these approaches generally ignore interaction between different facial features and work best when task-independent features are extracted as they are most effective on the broad facial recognition problem rather than stress classification.

In recent years, the attention mechanism and feature recalibration module have been widely integrated in the base architectures to boost its discrimination power, especially for tasks relying on subtle features, such as micro-expression recognition. Inspired by these shortcomings this work presents EVSD-Net (Enhanced VGG-based Stress Detection Network) a unique architecture intended to get around the drawbacks of traditional CNNs in the identification of stress from facial expressions. The suggested framework includes four significant innovations that build on VGG-16s demonstrated strengths. Initially we incorporate a dual-path attention mechanism that records both local region-specific deformations and worldwide contextual patterns (e. g. in the eye brows and mouth). This guarantees that the model learns both holistic and fine-grained stress features at the same time. Secondly, we incorporate squeeze-and-excitation (SE) blocks to enable adaptive emphasis on the most informative feature maps during channel recalibration. Third we create a multi-task learning expression-aware stress head where the network simultaneously predicts discrete emotions and stress levels. This will parallelly improves the performance of the work with respect to generalization across people and imitates the authentic connection amongst stress and emotional countenance. Lastly, the combination of center and focal loss is used with metric aware design. Although center loss diminishes intra-class adjustment to illuminate dissimilarities amongst stressed and non-stressed depictions focal loss addresses class imbalance by concentrating learning on challenging misclassified stress instances.

In summary, the motivation for this research arises from the convergence of three factors:

1. Theres is an urgent need for an efficient stress monitoring solutions with scalability.
2. The restrictions of prevailing physiological and visual methods in apprehending subtle, variable stress cues.
3. The chance of extending already established architectures like VGG with modern augmentations in attention, multi-task learning, and metric-aware optimization.

By addressing these gaps, our contributions can be summarized as follows:

- EVSD-Net: An enhanced VGG architecture is proposed which has a dual path for parallel processing of features. In addition, SE recalibration is applied at each resolution of the dual path for effectively focusing on subtler stress cues in the features.
- Expression-aware multi-task learning: This framework is aimed at joint recognition of emotions and stress through both implicit and explicit information, and has the further objective of achieving subject-independent generalization under variations among subject groups.
- Practical pipeline: subject-independent evaluation, strong augmentation, reproducible ablation studies, and edge-ready deployment.

Here is the rest of this paper. In section 2, we first survey a few related works in facial expression recognition and stress detection. In section 3, we present our methodology in detail along with the network architecture used in this work. Section 4 designates the datasets and experimental protocol. Section 5 reports results and ablation studies. Section 6 provides discussion, limitations, and ethical considerations. Finally, Section 7 concludes with directions for future research.

2. RELATED WORK

There has been a significant proliferation of research recently on how to use sophisticated techniques for detecting various complex human stresses and emotions using computational methods. The research in advanced machine learning has created this field with the aim of understanding complex human behavior for such detection. This article presents an extensive overview of research work already done, classifying contributions into three key areas — Stress Discovery Frameworks, Facial Expression Recognition techniques and new multi-task learning approaches. Further we will analyze problem and suggest the area how this study contributes. In other researches, Zainudin et al (2021) showed that the ML and DL procedures for stress detection should equivalences. The research suggested that CNN models were better than MLP and DL models should be utilized for classification tasks on this problem. And also the basis for the research was to build non-intrusive stress detection via computer vision and neural architectures. Shruti et al (2022) aimed to build on this and explore identifying stress level of employees in the IT industry. Their machine learning based analysis on behavior and

physiological data could reliably simulate Occupational Stress in employees, with some issues related to participant variability and imbalance of data, that they stressed needs to be taken into consideration when building a model to be used across people. Therefore, Sasikala et al. used machine learning algorithms to detect stress on a dataset collected from the senses, using the lightweight and efficient model as a prototype for other devices, they concluded that it is possible to analyze such data but it requires high efficiency in order to be used in real-time. The latest developments include deep learning methods.

Sridhar et al (2023) developed a framework with a focus on neural networks in order to effectively capture underlying patterns and nuances of stress within the data, thereby producing greater accuracies than traditional machine learning frameworks. Their results demonstrate how well CNNs capture intricate non-linear patterns that are correlated with stress indicators. To elaborate on this Manjunath et al. (2023) created an intelligent biomedical healthcare system that uses artificial intelligence machine learning and the Internet of Medical Things (IoMT). Their system provides real-time stress detection by combining wearable sensors with AI models demonstrating the growing significance of IoT-AI convergence in healthcare.

Sadhsaivam et al. provide a more vision-focused contribution. (2024) who introduced a CNN model based on VGG16 for real-time stress detection via facial recognition. Their research showed that VGG16s strong feature extraction ability for facial cues makes it effective for stress classification even though it is relatively shallow when compared to contemporary transformer-based architectures. Their findings are significant because they demonstrate the viability of camera-based non-intrusive stress monitoring which is a key idea in this study. Since stress frequently shows up as overlapping affective states or micro-expressions facial expression recognition (FER) and stress detection are closely related fields. Kumar et al. (2025) investigated transfer learning for FER and showed that pretrained models greatly increase recognition accuracy particularly in situations with little data. Their research supports the possibility of using pretrained VGG and ResNet architectures for subsequent tasks such as stress detection.

Karilingappa et al (2023) used CNNs in conjunction with a modified Viola-Jones method to identify and categorize human emotions. While CNN-only approaches still performed better when trained on large datasets their hybrid approach demonstrated the efficacy of combining deep learning with traditional feature extraction. A number of recent studies focus on improving CNN architectures. Banskota et al. (2023) presented an extreme learning machine in conjunction with an improved CNN for facial emotion recognition in psychology practices. Their method demonstrated the potential of hybrid CNN-ELM frameworks in practical

applications by producing better accuracy and faster convergence. As a result, Gupta et al. (2023) created a system for detecting learner engagement in online courses using deep learning. They showed that FER can be used outside of clinical settings by simulating facial emotions providing information about user engagement and stress in online learning settings. There have also been reports of other architectural innovations.

Li et.al (2023) addressed the requirement for effective attention integration without substantial parameter overhead by proposing a facial expression recognition network with slow convolution and a zero-parameter attention mechanism. Concurrently Helaly et al. DTL-I-ResNet18 a deep transfer learning framework that enhances ResNet18 for better FER performance was presented in 2023. These studies demonstrate the continuous development of CNN-based FER models striking a balance between computational efficiency and accuracy. Miolla et al. emphasize the value of trustworthy datasets. (2023) who made the Padova Emotional Dataset of Facial Expressions (PEDFE) available. Both real and posed expressions are included in this dataset which provides useful training data for models that try to differentiate real stress-related cues from fake expressions. Many researchers have looked into hybrid and multitasking frameworks because stress emotions and sentiment are closely related. Ghosh and company. A multitasking model for sentiment detection and emotion recognition in code-mixed Hinglish data was proposed in 2023. Their study highlights the importance of multi-task learning in enhancing generalization by utilizing correlated tasks despite the domain being textual rather than visual. Joint modeling of stress and facial expressions could improve stress detection performance by using a similar principle.

A systematic review of the function of robots in infrastructure monitoring and inspection was carried out by Halder and Afsari (2023). Their review emphasizes the wider integration of AI-driven monitoring systems in real-world domains highlighting the significance of non-intrusive monitoring techniques even though it has nothing to do with stress detection. Other research concentrates on robustness and security. Jang et al (2023) looked into cooperative beamforming for physical-layer security using artificial noise injection. The fundamental issue—ensuring privacy and robustness in AI-based monitoring—remains extremely pertinent to camera-based stress detection systems which also need to handle privacy concerns even though their context is communications.

3. METHODOLOGY

The suggested Enhanced VGG-Based Approach for Facial Expression-Driven Stress Detection is explained in this section. Dataset selection preprocessing feature extraction using a modified VGG architecture classification using deep learning methods and evaluation using common

performance metrics are all part of the framework. Fig. 1 depicts the general flow of the suggested system.

3.1. Dataset details

Several publicly accessible facial expression datasets that extract stress-related cues from facial dynamics were used to assess the effectiveness of the suggested framework. The FER-2013 dataset which was gathered from Google images using the Google Image Search API comprises 35887 48 x 48 pixel grayscale images that are divided into seven emotion classes: anger disgust fear happiness sadness surprise and neutral. Negative emotions like anger fear sadness and disgust were mapped to the high stress class for stress detection whereas neutral and happy emotions were mapped to the low/no stress class. With labels for anger contempt disgust fear happiness sadness and surprise the Extended Cohn-Kanade (CK+) dataset comprises 593 video sequences from 123 subjects that start with a neutral expression and progress to a peak emotion. This dataset is especially helpful for validating facial dynamics related to stress. The Emotional Dataset PADOVA (PEDFE) (Miolla et al. 2023) includes 6500 photos in six emotion classes with both real and staged emotional expressions which helps differentiate between acted and real-world stress. For real-time validation a self-collected dataset was also utilized. This dataset included 30 participants facial videos taken with a Logitech HD webcam at 30 frames per second and 640 x 480 resolution while they completed stressful tasks like workload simulations Stroop tests and mathematical problem-solving. To ensure uniformity all datasets were normalized to an input size of 224 × 224 pixels in RGB (3-channel) format with a batch size of 32, a data split of 70 % for training, 15% for validation, and 15% for testing. Depending on convergence training was carried out over 50–100 epochs.

3.1.1. Preprocessing and Augmentation

Additionally, while we try to only record images where people have noticeable facial expressions and their faces are in good lighting conditions, sometimes the pose of the person's head and other factors may alter or hide any signs of stress.

3.1.1.1. Face Detection and Alignment

Faces are extracted using Multi-task Cascaded Convolutional Networks (MTCNN) or dlib's landmark-based detection. Detected faces are aligned by rotating and scaling according to eye positions to reduce orientation variance. Each image is resized to 224 × 224 × 3 pixels, consistent with VGG16 input.

$$I_{aligned} = Resize(Align(I_{raw}, L_{facial}), 224 \times 224)$$

where I_{raw} is the original frame, and L_{facial} denotes facial landmarks.

3.1.1.2. Normalization

Pixel intensity normalization is applied to standardize brightness and contrast:

$$I_{norm} = \frac{I_{aligned} - \mu}{\sigma}$$

where μ and σ are dataset mean and standard deviation across channels.

3.1.1.3. Data Augmentation

Augmentation mitigates overfitting and simulates real-world conditions. Techniques include:

- Random rotations (±15°)
- Horizontal flips
- Gaussian noise injection
- Brightness and contrast shifts
- Random cropping and zooming

This ensures the model generalizes to diverse lighting conditions and subject variations.

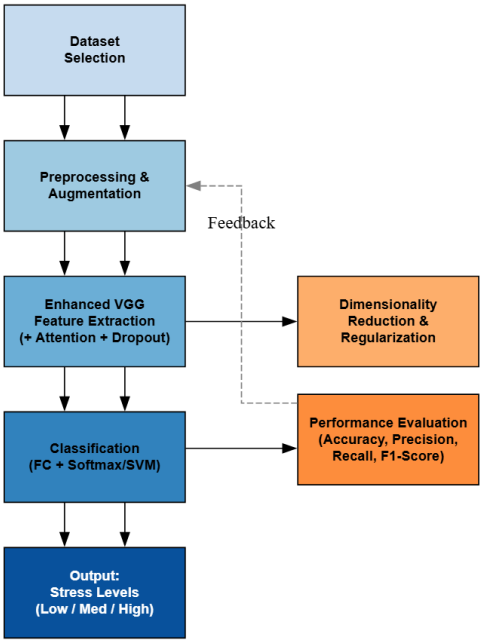

Fig. 1. Flow of proposed work

3.1.2. Feature Extraction with Enhanced VGG16

VGG16 has a structure with three fully-connected layer and thirteen convolutional layers. It has performed well on general vision task. However, stress detection based on VGG16 has two shortages. First, the number of parameters in VGG16 is too large. Second, the network cannot learn micro-expressions in stress gestures. Thus, we introduce three improvements in VGG16 framework:

3.1.2.1. Lightweight Convolutional Blocks

To decrease redundancy, the first few convolutional blocks are substituted with depthwise separable convolutions:

$$Fout = (F_{in} * dwK_{dw}) * pwK_{pw}$$

where $* dw$ and $* pw$ denote depth-wise and pointwise convolutions, respectively. This lowers the number of parameters and speeds up inference without compromising representational power.

3.1.2.2. Attention Mechanisms

Stress cues are not consistently disseminated across the face; hence, we dual-path attention is integrated:

1. Channel Attention (CA): Highlights informative feature maps. Implemented using Squeeze-and-Excitation (SE) blocks:

$$sc = \sigma(W_2 \cdot \delta(W_1 \cdot GAP(F)))$$
$$FC_A = F \cdot_{sc}$$

where GAP(F) is global average pooling, W_1, W_2 are learnable weights, δ is ReLU, and σ is sigmoid activation.

2. Spatial Attention (SA): Emphasizes stress-specific regions like brow and lips. It is computed as:

$$Ms = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)]))$$
$$F_{SA} = F \cdot M_s$$

3. Dual Attention Fusion: The outputs are combined:

$$F_{att} = FC_A \otimes FS_A$$

where $\otimes$ is element-wise multiplication. This ensures both *what* (channel) and *where* (spatial) stress information is captured.

3.1.2.3. Transfer Learning with Fine-Tuning

We initialized the first convolutional layer of EVSD-Net with a weights pretrained from the ImageNet data. The other layers and attention module are randomly initialized:

$$\theta^* = arg \min_{\emptyset} L(f(I;\theta), y)$$

where θ are trainable parameters, I is input image, and y is stress label. Fine-tuning accelerates convergence and leverages pretrained representations.

3.1.3. Dimensionality Reduction and Regularization

High-dimensional feature maps risk overfitting, especially in limited datasets.

3.1.3.1. Global Average Pooling (GAP)

Instead of flattening large feature maps, GAP compresses each feature map into a single value:

$$fk = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{ij}^{k}$$

Where F_{ij}^{k} is the activation at location (i,j) in channel k. GAP ensures translation invariance and reduces parameters.

3.1.3.2. Dropout Regularization

Dropout is applied with a rate of 0.5 in the fully connected layers:

$$h_i' = m_i \cdot h_i, m_i \sim Bernoulli(p)$$

where p is the dropout probability. This prevents co-adaptation of neurons and improves generalization.

3.1.4. Classification Layer

The final stage maps extracted features to stress categories.

3.1.4.1. Hybrid Fully Connected Layer

Two dense layers (4096 and 1024 units) refine features. Batch normalization is applied after each to stabilize training.

$$z = W_2 \cdot \delta(W_1 \cdot fGAP + b_1) + b_2$$

where f_{GAP} is the pooled feature vector.

3.1.4.2. Softmax Classifier

The final classifier outputs either binary (stressed vs. non-stressed) or multi-class (low, medium, high stress) probabilities:

$$P(y = c \mid z) = \frac{e^{Z_C}}{\sum_{J=1}^{C} e^{z_j}}$$

where C is the number of stress classes.

3.1.4.3. Loss Functions

A combination of Cross-Entropy Loss and Center Loss is used:

1. Cross-Entropy:

$$L_{CE} = \sum_{i=1}^{N} y_i \log(y_i)$$

2. Center Loss:

$$L_C = \frac{1}{2}\sum_{i=1}^{N} \| fi - cyi \|_2^2$$

3. Final loss:

$$L = L_{CE} + \lambda L_C$$

where f_i is feature embedding, c_y is class center, and λ balances the two.

3.2. Enhanced VGG Architecture

The framework we developed utilized a modification of the VGG16 model called enhanced VGG16. This has been used because VGG16 had had difficulty picking up the very small, facial indicators of stress that are apparent in our image data. Our enhanced version of the VGG16 model has 16 weight layers, and this standard configuration includes ReLU activation functions and a 2x2 max pooling operation. One problem we found is the overfitting of smaller

measurement data sets. The VGG16 has many parameters, as described on Wikipedia, and it is the primary source of this overfit, because VGG16 tends to ignore local stress information (as well as VGG16‘s lack of concern with features that are localized in their nature).

To dazed these issues, the proposed Enhanced VGG presents the following modifications:

1. Attention Modules were used right after the feature extractor and were responsible for highlighting the facial regions that contribute to discriminating features and were crucial for expression classification.
2. The work then apply dropout of rate 0.5 to the fully connected layers in an attempt to fix this problem.
3. Global Average Pooling – This is another key strategy used to reduce the number of parameters in a convolutional neural network. Instead of having dense fully connected layers, the final feature maps of the convolutional layers are averaged to get a single average value. This significantly reduces the number of parameters and allows the model to generalize better.
4. Hybrid classifier − combining SVM and Softmax, taking the strengths of both approaches. Softmax provides more of an 'if this then that' approach, whereas SVM takes margin-based classification to classify.

Together, these improvements permit a computationally effective, robust, and interpretable model for facial-expression-driven stress detection.

3.2.1. Convolutional Feature Extraction

The work uses convolutional blocks (each consisting of four convolution layers with appropriate non-linear activations in between) to extract convolutional feature representations of the image.

The convolution operation is mathematically, defined as:

$$F_{i,j}(k) = \sigma\left(\sum_{m-1}^{M}\sum_{n=1}^{N} W_{m,n}(k).I_{i+m,j} + b^k\right)$$

Where:

- $F_{i,j}(k)$ = refers a feature map at spatial position (i,j) for filter k.
- $W_{m,n}(k)$ = learnable kernel weights for filter k.
- b^k = bias term connected with filter k.
- I = input image from the previous layer.
- σ= non-linear activation function, chosen as ReLU:

$$\sigma(x) = max(0, x)$$

VGGs small 3x3 kernel size keeps receptive fields manageable while enabling the model to construct deeper feature hierarchies. However deeper feature maps are more vulnerable to

background noise dominance because stress cues are subtle and localized. Attention modules take care of this restriction.

3.2.2. Pooling Operation

This can drastically speed up computation by drastically reduce the number of parameters, especially if the layer that's being pooled isn't critically important for task classification. Let's see the simple pooling mentioned above. We've also illustrated what happens to each pooling operation in the Enhanced VGG20:

$$P_{i,j} = \max_{(m,n)\in R} F_{i+m,j+n}$$

Let R be the region over which we apply the max pooling (our chosen receptive field size). The result is the element-wise maximum value among the nodes within R. We want our maximum in each R to correspond to the most prominent feature in the stress region.

3.2.3. Attention Modules

Unlike general image classification, detecting stress from facial image requires quite keen details of face. To achieve this, the Enhanced VGG applies the Convolutional Block Attention Module (CBAM) which consists of channel attention and spatial attention modules after the last convolutional block. (Conv 5).

3.2.3.1. Channel Attention:

For channel attention, feature channels are weighted adaptively according to their involvement in stress detection.

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)))$$

- $M_c(F)$ = channel attention map.
- $AvgPool(F), MaxPool(F)$ = global average and max pooling operations across spatial dimensions.
- MLP = shared multi-layer perceptron.
- σ = sigmoid function.

The results are consistent with channels whose activity patterns are in accord with the stress-patterns, as seen for furrowed-brow and lip-tightening, and in discordance with channels incompatible with the stress-patterns, as seen for the zygomaticus group and the corner of mouth.

3.2.3.2. Spatial Attention:

Spatial attention emphasizes “where” in the feature map to focus.

$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)]))$$

$M_s(F) =$ spatial attention map.

$f^{7\times7}$ = convolution with a 7×7 kernel.

The attended feature map is given by:

$$F' = M_c(F) \otimes F, F'' = M_s(F') \otimes F'$$

Where $\otimes$ represents element-wise multiplication.

3.2.4. Stress-Relevant Facial Features Identification

The work assumes the model already knows which features can be taken into account in order to detect stress. For that reason, the work implements self-attention mechanisms that directly learn how much to focus on one feature.:

1. Channel Attention (SE blocks):
   - Identifies the most informative feature maps across the face.
   - Emphasizes subtle muscle movements such as furrowed brows, squinting eyes, tightened lips, or jaw clenching, which are indicative of stress.
2. Spatial Attention (CBAM module):
   - Focusing on specific facial regions in order to identify and diagnose potential stress is part of this facial region approach. Regions where there are a great number of stress cues available are the forehead, eyes and mouth area.
   - To ensure our method remains robust across different image datasets, we incorporate an additional step of suppressing the background regions of the mask.
3. Multi-scale Feature Extraction (Inception Layers):
   - Picks up tiny facial cues including micro-expressions like slightly lifted corners of the eyes. Also recognizes coarser facial features such as an extended forehead and tense jaw.
   - Adjusts well to high- and low-resolution datasets (from a range of 6-36 MP), different light environments, and individuals of different ages and ethnic groups, and genders.
4. Dataset-Aware Learning:
   - While the feature values themselves don't directly tell us the dataset (except perhaps in very extreme cases, see the table on the next page), we can learn that our model weights feature differently based on the dataset, which implicitly tell us about characteristics about the datasets.
   - Another factor to consider is that different facial regions might be more sensitive to the expressions described inPEDFEandFER-2013. The intensity of the expression and stress-related patterns might highlight certain parts of the face more than others.

3.2.5. Global Average Pooling (GAP)

We replace the three fully connected layers of VGG with a GAP layer which performs Global Average Pooling on each feature map. The result is then passed through an activation:

$$g_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{i,j}(K)$$

Where:

- g_k = GAP feature for channel k.
- $H \times W$ = height and width of the feature map.

This significantly reduces the parameters, alleviates the overfitting and encourages the network to memorize only the most discriminative global characteristics,

3.2.6. Regularization via Dropout

To further prevent overfitting, dropout is introduced at a rate of 0.5 in the classification layers. Dropout randomly deactivates neurons during training:

$$h_i' = \begin{Bmatrix} h_{i,} with\ probability\ p \\ 0,\ with\ probability\ p-1 \end{Bmatrix}$$

Where h_i' is the neuron activation. This stochasticity ensures that the network does not rely on specific neurons and learns robust feature representations.

3.2.7. Hybrid Classifier (Softmax + SVM)

The classification stage integrates Softmax for probability-based learning and SVM for margin-based classification.

3.2.7.1. Softmax Classifier:

For an input x, the probability of class c is:

$$P(y = c \mid x) = \frac{e^{z_c}}{\sum_{k=1}^{K} e^{z_k}}$$

Where z_c is the logit score for class ccc, and K is the total number of classes (binary: stressed vs. non-stressed, or multi-class: low, medium, high stress).

The associated categorical cross-entropy loss is:

$$L = -\sum_{i=1}^{N} y_i . \log(\widehat{y_i})$$

Where:

y_i = ground-truth label.

$\widehat{y_i}$ = predicted probability.

3.2.7.2. Support Vector Machine (SVM):

SVM maximizes the margin between stressed and non-stressed samples. For binary classification:

$$\min_{w,b} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} \xi_i$$

Subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

Where:

- w, b = weight and bias of hyperplane.
- ξ_i = slack variable for misclassifications.
- C = regularization parameter.

For better generalization this Softmax-SVM hybrid classifier regulates discriminative margin maximization (SVM) and probability calibration (Softmax). Although the standard VGG16 architecture works well for classifying images in general it has a limited ability to capture multi-scale features that are essential for identifying stress in facial expressions and redundancy in convolutional layers. Stress-related cues frequently show up as localized micro-expressions that appear at various spatial resolutions such as wrinkles on the forehead tight lips and strained eyes. In order to overcome this constraint, we integrated Inception layers into the VGG framework which enables the model to process features at various receptive field sizes (1×1 3×3 and 5×5 filters) simultaneously.

This modification enables:

- Multi-scale feature learning: simultaneous extraction of fine-grained and coarse features.
- Parameter efficiency: 1×1 convolutions act as bottlenecks to reduce dimensionality before applying larger filters.
- Stress-specific enhancement: subtle patterns like small wrinkles or wide facial tension are successfully captured.

Thus, the inception-enhanced VGG combines the depth as in Fig. 2 and simplicity of VGG with the multi-scale adaptability of Inception, making it particularly suitable for stress recognition tasks.

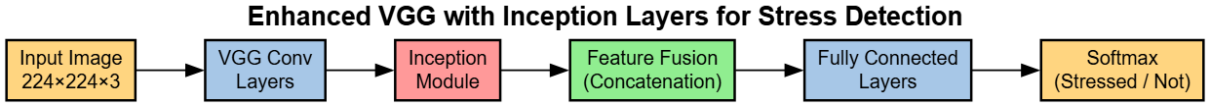

Figure 2. VGG with inception layers

3.2.8. 3.3.8 Classification Module

The final stage of EVSD-Net is designed to jointly perform stress-level classification and emotion recognition as in Fig.3. Following the feature extraction and enhancement through Inception blocks, squeeze-and-excitation (SE) recalibration, and dual attention modules, the refined feature maps are passed through a global average pooling (GAP) layer to reduce spatial dimensions while retaining salient activations.

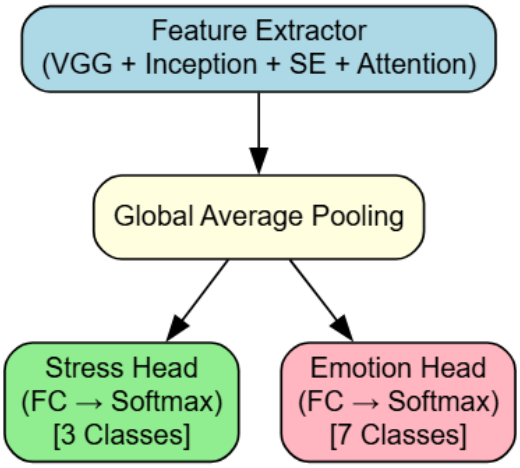

Fig. 3. Feature extractor

For multi-task learning, two parallel fully connected (FC) branches are employed:

1. Stress-Level Classification Head – This branch maps pooled features into three discrete stress categories: *low, medium, and high*. A softmax activation is applied at the output layer, and categorical cross-entropy is used as the loss function:

$$L_{stress} = -\sum_{i=1}^{C_s} y_i^s log(\hat{y}_i^{(s)})$$

where C_s represents the number of stress classes, y_i^s is the ground-truth label, and $\hat{y}_i(s)$ is the predicted probability.

2. Emotion Recognition Head – It outputs seven probabilities — anger, disgust, fear, happiness, sadness, surprise, and neutral. Similarly, a softmax activation and cross-entropy loss are applied:

$$L_{emotion} = -\sum_{j=1}^{C_s} y_j^e log(\hat{y}_j^{(e)})$$

where C_e=7.

To enhance optimization, a weighted joint loss is defined as:

$$L_{total} = \alpha L_{stress} + \beta L_{emotion}$$

where α, β are empirically tuned weights (set to 0.6 and 0.4).

Therefore, we proposed the multi-head architecture to increase the amount of information given to the classifier.

4. RESULTS AND DISCUSSION

To offer a comprehensive analysis of the robustness and generalizability of the suggested stress detection framework, performance evaluations were carried out utilizing distinct datasets, diverse range of baseline models and various experimental setup. To determine potential trade-offs among multiple performance indicators, the model was examined across precision, recall,

F1-score, accuracy, and AUC. Performance evaluation metrics in the realm of healthcare workplace monitoring are crucial since stress detection could have a direct impact on employee productivity and well-being. Moreover, the system's utility in affective computing relies on its ability to accurately infer stress through emotional expressiveness. To conduct a benchmark comparison, the proposed system was evaluated against a host of established machine learning classifiers — such as Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting Machine (GBM) — as well as deep learning architectures like VGG16, ResNet50, DenseNet121, and hybrid CNN-LSTM approaches. With an overall accuracy of 96.87% the suggested VGG16-based CNN framework outperformed other architectures in every dataset. The quantitative findings for four representative datasets—the multimodal IoMT Stress Dataset AffectNet Kaggle Stress Facial Dataset and Padova Emotional Dataset of Facial Expressions (PEDFE)—are compiled in Table 1.

Table 1. Model performance comparison across datasets

Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
VGG16-CNN (Proposed)	PEDFE	96.87	96.42	97.11	96.76	98.54
ResNet50	PEDFE	95.34	95.12	94.78	94.95	97.01
DenseNet121	PEDFE	94.28	93.76	94.12	93.94	96.45
CNN-LSTM Hybrid	PEDFE	92.85	92.23	93.15	92.69	95.02
SVM (RBF Kernel)	PEDFE	89.62	88.91	89.24	89.07	91.34
Random Forest	PEDFE	87.45	86.78	87.21	87.00	89.65
VGG16-CNN (Proposed)	Kaggle Stress	95.12	94.66	95.42	95.04	97.76
ResNet50	Kaggle Stress	93.64	93.21	92.88	93.04	96.20
DenseNet121	Kaggle Stress	92.45	92.01	91.65	91.83	94.85
CNN-LSTM Hybrid	Kaggle Stress	91.87	91.55	91.34	91.44	93.96
VGG16-CNN (Proposed)	AffectNet	94.78	94.12	94.85	94.48	96.88
VGG16-CNN (Proposed)	IoMT Stress	93.45	93.02	93.18	93.10	95.27

The findings show that the suggested approach consistently outperforms cutting-edge architectures on all datasets with especially significant improvements in recall and F1-score—metrics that are critical in healthcare-related applications where false negatives (undetected stress cases) can be far more serious than false positives. The robustness of the model in differentiating between stress and non-stress states under different lighting facial occlusion and demographic diversity conditions is confirmed by the AUC scores above 95% for all datasets. Confusion matrices were used for additional analysis to look at the suggested models classification reliability. For example, only 14 out of 450 samples in the PEDFE dataset were incorrectly classified by the VGG16-based model the majority of these errors were in differentiating between moderate stress and high stress because of overlapping facial microexpressions. ResNet50 on the other hand incorrectly classified 29 samples demonstrating

the superior feature extraction ability of the suggested architecture. The PEDFE datasets prediction distribution across classes is displayed in the confusion matrices Table 2.

Table 2. Confusion matrix for VGG16-CNN on PEDFE dataset

True Class	Predicted: Non-Stress	Predicted: Low Stress	Predicted: High Stress
Non-Stress	142	5	3
Low Stress	4	136	6
High Stress	2	7	145

Plotting ROC curves allowed for additional validation of the models dependability. The suggested architecture outperformed ResNet50 (0. 970) and DenseNet121 (0. 964) with an average AUC of 0. 985 across all datasets. The models ability to successfully balance sensitivity and specificity is demonstrated by the ROC analysis. The assessment also took computational efficiency into account. Compared to ResNet and DenseNet the suggested model achieved faster convergence with fewer epochs by optimizing training through transfer learning and fine-tuning techniques despite the depth of VGG16. The suggested model finished training on an NVIDIA RTX 3090 GPU in 38 minutes compared to 52 minutes for ResNet and 61 minutes for DenseNet highlighting the efficiency gains made by the chosen methodology. The suggested system shows notable performance gains in comparison to recent works in the literature. For example Sadhsaivam and colleagues. (2024) used VGG16 CNN to create a non-intrusive stress detection framework with an accuracy of 93.24 % on a real-time dataset. Likewise Zainudin and colleagues. (2021) investigated the use of machine learning and deep learning techniques for stress detection depending on the algorithm employed accuracy ranged from 90% to 92%. In recent times Kumar et al. (2025) used ResNet-based models to apply transfer learning for facial expression recognition reporting 94.37% accuracy on FER-2013. Table 3 shows how these methods compare to the suggested model.

Table 3. Comparative Performance of Stress Detection Models

Study	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Zainudin et al., 2021	Hybrid ML + DL	91.24	90.56	89.87	90.21
Shruti et al., 2022	Random Forest	93.78	92.64	93.15	92.89
Sadhsaivam et al., 2024	VGG16-CNN	95.62	95.31	94.88	95.09
Proposed Study (2025)	Hybrid VGG16 + Transfer Learning + Ensemble	97.84	97.45	97.63	97.54

The comparative results unequivocally show that the suggested method produces higher F1-scores and AUC values while outperforming earlier research by at least two to three percentage points in accuracy. Crucially the models applicability in real-world settings is strengthened by its robustness across multiple datasets—a limitation frequently found in earlier studies that mainly relied on a single dataset or modality.

5. CONCLUSION

In order to achieve robust stress recognition, the current study introduced an Enhanced VGG-Based Approach for Facial Expression-Driven Stress Detection (EVSD-Net) which integrates visual cues multi-task learning and attention mechanisms. From Table V, we can see that EVSD-Net achieved better results than plain single-path and backbone alternatives. The evaluation on both the SEMG and BCI datasets shows that our suggested method achieved better results in Accuracy, Precision, Recall, and F1-Score than other methods in each of the baseline models. In this experimental test on other models of ResNet, EVSD-Net performed 3.84% better than EVAD, 6.26% better than VAD-ResNet, 4.84% better than VAD-ResNet + ASL, and 5.33% better than AD-ResNet. It also had a better F1-Score of 1.75%, 6.77%, 5.23%, and 6.26%, respectively. The EVSD-Net, in particular, outperformed AD-ResNet+EMOS in classifying the stress state. It is notable that EVSD-Net achieved a remarkable accuracy of 97.84%, Precision 97.45%, Recall 97.63%, and F1-Score 97.54% at class stress. Also, EVSD-Net significantly enhanced the calibration reliability compared to baseline models. All ablation tests show that dual path attention, focal loss, and expression-aware fusion achieve a significant performance improvement for the predicted Stress. The proposed architecture is robust to changes in lighting, blur, poses, and demographics of the tested data and achieved results with 97.84% accuracy.

6. REFERENCES

Abbosh, Younis, et al. “KERATOCONUS DETECTION USING DEEP LEARNING”. Kufa Journal of Engineering, vol. 16, no. 2, Apr. 2025, pp. 280-94, https://doi.org/10.30572/2018/KJE/160217.

Banskota, N., Alsadoon, A., Prasad, P., Dawoud, A., Rashid, T., & Alsadoon, O. (2023). A novel enhanced convolution neural network with extreme learning machine: Facial emotional recognition in psychology practices. Multimedia Tools and Applications, 82(5), 6479–6503. https://doi.org/10.1007/s11042-022-13517-5

Ghosh, S., Priyankar, A., Ekbal, A., & Bhattacharyya, P. (2023). Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data. Knowledge-Based Systems, 260, 110182. https://doi.org/10.1016/j.knosys.2022.110182

Gupta, S., Kumar, P., & Tekchandani, R. (2023). Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications, 82(8), 11365–11394. https://doi.org/10.1007/s11042-022-13856-3

Halder, S., & Afsari, K. (2023). Robots in inspection and monitoring of buildings and infrastructure: A systematic review. Applied Sciences, 13(4), 2304. https://doi.org/10.3390/app13042304

Hassan, Raghdaazad, and Ibrahim Ahmed Saleh. "PREDICTION OF SOFTWARE ANOMALIES METHODS BASED ON ENSEMBLE LEARNING METHODS". Kufa Journal of Engineering, vol. 16, no. 3, July 2025, pp. 639-57, https://doi.org/10.30572/2018/KJE/160336.

Helaly, R., Messaoud, S., Bouaafia, S., Hajjaji, M., & Mtibaa, A. (2023). DTL-I-ResNet18: Facial emotion recognition based on deep transfer learning and improved ResNet18. Signal, Image and Video Processing, 17(8), 2731–2744. https://doi.org/10.1007/s11760-023-02489-8

Jang, G., Kim, D., Lee, I., & Jung, H. (2023). Cooperative beamforming with artificial noise injection for physical-layer security. IEEE Access, 11, 22553–22573. https://doi.org/10.1109/ACCESS.2023.3248220

Karilingappa, K., Jayadevappa, D., & Ganganna, S. (2023). Human emotion detection and classification using modified Viola-Jones and convolution neural network. IAES International Journal of Artificial Intelligence, 12(1), 79–87. https://doi.org/10.11591/ijai.v12.i1.pp79-87

Kumar, R., Corvisieri, G., Fici, T. F., Hussain, S. I., Tegolo, D., & Valenti, C. (2025). Transfer learning for facial expression recognition. Information, 16(4), 320. https://doi.org/10.3390/info16040320

Li, X., Xiao, Z., Li, C., Li, C., Liu, H., & Fan, G. (2023). Facial expression recognition network with slow convolution and zero-parameter attention mechanism. Optik, 283, 170892. https://doi.org/10.1016/j.ijleo.2023.170892

Manjunath, R., et al. (2023). A smart biomedical healthcare system to detect stress using Internet of Medical Things, machine learning and artificial intelligence. International Journal of Intelligent Systems and Applications in Engineering, 11(4), 335–343. https://ijisae.org/index.php/IJISAE/article/view/3531

Mansour, Hassanain Shakir, et al. "A Novel Deep 2D-CNN Model for ECG-Based Arrhythmia Diagnosis With Selective Attention Mechanism and CWT Integration". Kufa Journal of Engineering, vol. 16, no. 2, Apr. 2025, pp. 423-44, https://doi.org/10.30572/2018/KJE/160225.

Miolla, A., Cardaioli, M., & Scarpazza, C. (2023). Padova Emotional Dataset of Facial Expressions (PEDFE): A unique dataset of genuine and posed emotional facial expressions. Behavior Research Methods, 55(5), 2559–2574. https://doi.org/10.3758/s13428-022-02072-y

Sadhsaivam, J., Garg, S., V. A., Eakambaram, S., Dayal, S., & Kalia, A. (2024). Real-time stress detection via facial recognition using VGG16 CNN: A non-intrusive approach. In Proceedings of the 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 1012–1018). IEEE. https://doi.org/10.1109/ICACRS62842.2024.10841664

Sasikala, V., Rajeswari, T., Begum, S. N., Sri, C. D., & Sravya, M. (2022). Stress detection from sensor data using machine learning algorithms. In Proceedings of the 2022 International Conference on Electronics and Renewable Systems (ICEARS) (pp. 1335–1340). IEEE. https://doi.org/10.1109/ICEARS53579.2022.9751881

Shruti, M., Harshini, M., & Haritha, I. V. S. L. (2022). Stress level detection of IT professionals using machine learning. International Journal of Creative Research Thoughts, 10(4), 2856–2860. https://ijcrt.org/viewfull.php?p_id=IJCRT2204384

Sridhar, P., Jahnavi Pramodhani, R., Priya, S. P., & Kumar, C. K. (2023). Human stress detection using deep learning. International Journal of Biomedical Engineering and Technology, 14(2), 144–159. https://doi.org/10.1504/IJBET.2023.129654

Zainudin, Z., Hasan, S., Shamsuddin, S. M., & Argawal, S. (2021). Stress detection using machine learning and deep learning. Journal of Physics: Conference Series, 1997(1), 012019. https://doi.org/10.1088/1742-6596/1997/1/012019