Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 9, No. 1, January 2027, pp: 131-155 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

An Interpretable Lightweight CNN Framework with Grad-CAM for Stress-Related Facial-Expression Classification

Sathish G[1], Priya Matta [2], Anusha N[3], P. Pattabhirama Mohan[4], Aanandha Saravanan K[5] and Dilli Ganesh V[6].

[1] Department of Computer Science, Bharath Institute of Science and Technology, BIST, Chennai, India
[2] Department of Computer Applications, Tula's Institute, Dehradun, India
[3] Nitte (Deemed to be University), NMAM Institute of Technology (NMAMIT), Dept of ISE, Nitte, Karnataka, India
[4] Department of Information Technology, Aditya University, Andhra Pradesh, India
[5] Department of ECE, Veltech Rangarajan Dr.Sagunthala R&D Institute of Science and technology, Chennai, India
[6] Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences-(SIMATS), Chennai, India

Corresponding author: Priya Matta. (e-mail: Dehradun Mattapriya21@gmail.com), **Author(s) Email**: Sathish G (e-mail: sathish.cse@bharathuniv.ac.in), Anusha N.(e-mail: anusha24@nitte.edu.in) P. Pattabhirama Mohan (e-mail: ramspatnala@gmail.com), Aanandha Saravanan K(e-mail: aanandhasaravanan@veltech.edu.in) and Dilli Ganesh V (e-mail: dilliganesh001@gmail.com)

Abstract

Background: Mental stress is a major public health concern associated with impaired cognitive performance, emotional well-being, and productivity. Conventional stress assessment methods, including self-report questionnaires and clinical observations, are subjective, intermittent, and unsuitable for continuous monitoring.

Aims: This study aims to develop and evaluate a lightweight, explainable deep learning framework for the noninvasive binary classification of stress-related and non-stress facial expressions using publicly available datasets.

Methods: Facial images from the FER2013 and Extended Cohn–Kanade (CK+) benchmark datasets were preprocessed through face extraction, resizing, normalization, and data augmentation. The FER2013 dataset was divided into training (70%), validation (15%), and testing (15%) sets, whereas the CK+ dataset was partitioned at the subject level into training (80%), validation (10%), and testing (10%) sets and evaluated using leave-one-subject-out (LOSO) cross-validation to prevent identity leakage. A lightweight convolutional neural network (CNN) with approximately 0.42 million parameters was trained to distinguish stress-related from non-stress facial expressions. Gradient-weighted Class Activation Mapping (Grad-CAM) was integrated to visualize the facial regions contributing to the model's predictions.

Results: On the FER2013 test set, the proposed CNN achieved an accuracy of 89.6%, an F1-score of 89.6%, and an area under the receiver operating characteristic curve (AUC) of 0.93. Under subject-independent LOSO cross-validation on CK+, the model achieved an accuracy of 89.8 ± 2.7% and an F1-score of 90.2 ± 2.6%. Compared with VGG16, ResNet50, and EfficientNet-B0 under identical experimental conditions, the proposed model demonstrated a favorable balance between predictive performance and computational efficiency, achieving an inference latency of 12.0 ± 0.8 ms on an NVIDIA RTX 4060 GPU.

Conclusion: This study presents a lightweight, interpretable, and reproducible deep learning framework that supports computationally efficient affective analysis and preliminary monitoring of stress-related facial expressions. However, because the binary classification labels were derived from emotion annotations rather than clinically validated stress measurements, the framework should be regarded as a stress-related proxy classification model and should not be deployed in clinical settings without further validation using clinically annotated datasets.

Keywords Stress Detection; Facial Expression Analysis; Convolutional Neural Network; Explainable Artificial Intelligence; Grad-CAM; Emotion Recognition; Mental Health Monitoring; Deep Learning

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

I. Introduction

In recent years, mental stress has become a major public health concern, where its effects have influenced the cognitive processing, the way one feels and the general wellbeing. Stress over a long period is linked to dire health issues as anxiety disorders, depression and cardiovascular diseases, hence the need to have early and accurate detection measures. There is a tendency of bias, subjectivity, and fastness that is found in the conventional stress assessment methods such as self-reporting and clinical assessments. As a result, the increased need in automated, objective, and real-time stress-related facial-expression classification systems is emerging that would help in the early diagnosis and intervention [1], [2]. The modern achievements in the field of artificial intelligence (AI), and in specific, deep learning, have brought a revolution in the industry of mental health monitoring as it is now possible to create intelligent systems that can process detailed patterns of human behavior [3].

Convolutional neural networks (CNNs) are deep learning models that have demonstrated strong performance in image-based tasks, such as recognition of facial expressions and the classification of emotions. Those methods exploit hierarchical feature extraction to detect the slight dissimilarities in facial expressions, which is suggestive of underlying emotional conditions [4], [5], [6]. Based on this, the analysis of facial expressions has emerged as a promising noninvasive process of identifying psychological states like stress and depression. Facial expressions are also a good way to learn about the human emotions because they are a physiological reaction to inner mental conditions which is not always under control. The recent research has investigated the possibility of using deep learning models to identify stress based on facial features in captured images or video streams. These systems use big datasets like AffectNet and FER2013 to train the powerful models that can identify the emotional states with high precision [7], [8], [9].

Moreover, recognition systems of facial expressions have been enhanced with the introduction of advanced architectures such as attention systems and hybrid models, which has enhanced the accuracy and generalization of the system [10], [11]. The lack of interpretability is the greatest limitation of current deep learning-based stress-related facial-expression classification systems despite considerable progress. Most models are black box-based, meaning they give predictions without describing the reasoning behind them and this is why they are not being adopted across sensitive areas like medical care. To overcome this problem, explainable Artificial Intelligence (XAI) methods have been brought into the light to provide information about the model decisions, which can be more transparent, trustworthy, and easy to use. Visualization of the key parts of the face that are relevant to the detection of stress can be achieved with the help of such methods as Grad-CAM and SHAP, which makes AI systems more interpretable and clinically relevant [12], [13], [14].

Although multimodal approaches combining facial expressions with EEG, speech, or wearable-sensor data can improve classification performance, they generally require specialized hardware and complex data-acquisition procedures [15]–[17]. Existing facial-expression-based approaches are easier to deploy, but many employ computationally intensive architectures, provide limited interpretability, or inadequately control identity leakage during evaluation. Furthermore, emotion-derived labels are sometimes presented as direct indicators of mental stress despite the absence of clinically validated stress measurements. Therefore, a clear research gap exists for a lightweight, explainable, and rigorously evaluated framework that explicitly treats emotion-derived labels as proxies for stress-related facial expressions rather than as clinical stress diagnoses.

To address this gap, the present study proposes a lightweight CNN containing approximately 0.42 million trainable parameters and integrates Grad-CAM as a post-hoc interpretability module. The framework uses FER2013 and CK+ datasets, applies identity-conscious evaluation procedures, and compares the proposed model with VGG16, ResNet50, and EfficientNet-B0 under consistent training conditions. The aim of this study is to develop and evaluate an efficient and interpretable framework for binary classification of stress-related and non-stress facial expressions using publicly available benchmark datasets. The principal contributions of this study are: (1) an explicit formulation of the task as stress-related proxy classification rather than clinically validated stress detection; (2) a compact CNN architecture suitable for low-latency analysis; (3) Grad-CAM-based visualization of prediction-relevant facial regions; (4) evaluation procedures designed to reduce identity leakage; and (5) comprehensive assessment using classification, ROC-AUC, computational-efficiency, ablation, baseline-comparison, and interpretability analyses.

II. Related Work

Recent breakthroughs in the area of artificial intelligence have made a substantial contribution to the conception of the automated stress and mental health analysis systems. The identification of stress and emotional states on the basis of various modalities such as facial expressions, physiological expressions, and multimodal data sources have been studied with the help of deep learning-based methods. These are the methods of giving the accurate, noninvasive, and

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

real-time so1lutions of mental health monitoring. There is also a significant amount of literature dedicated to the implementation of stress recognition based on deep learning algorithms on the facial expression recognition. Convolutional Neural Networks (CNNs) have gained immense use because of its ability to automatically extract hierarchical features of facial images. As it has already been proved by various studies, CNN-based models are effective to classify emotional states and identify stress-related patterns with benchmark datasets like AffectNet and FER2013 [18], [19], [20].

Moreover, more sophisticated architectures with the use of attention systems, as well as hybrid feature extractors have led to better classification accuracy and robustness even in general real-world conditions. Simultaneously, explainable artificial intelligence (XAI) has become popular within recent years to overcome the interpretability issue that deep learning models have. Conventional deep learning systems can be black-box models, and thus they cannot be used in high-stakes applications, like in healthcare. To address this drawback, scientists have proposed explanatory models that indicate and depict the most significant facial areas that make predictions. Grad-CAM, feature attribution methods, and others are all successful techniques used to increase the transparency and trust of a stress-related facial-expression classification system [21], [22].

In addition to the methods based on facial expressions, a number of studies have examined multimodal methods, which integrate facial data with physiological evidence of electroencephalography (EEG), speech and wearable sensor data. The needs of these systems are to enhance accuracy in detection by using complementary information across various sources. To illustrate, the emotion recognition models based on EEG and the hybrid systems that combine the data of eye movement and speech have demonstrated good performance in recognizing the

Table 1. Comparative summary of previous studies on facial-expression analysis and stress-related classification, including methods, data modalities, principal contributions, and reported limitations

Ref.	Method / Model	Data Type	Key Contribution	Limitation
[1], [6]	Ensemble CNN + XAI, YOLO-based models	Multimodal / Facial	Real-time stress detection with explainability	High computational complexity
[4], [28]	XAI + Deep Learning (CNN/Ensemble)	Facial expressions	Improved emotion recognition with feature importance	Limited generalization
[2], [3], [8]	AI-based mental health systems	General / Multimodal	Broad applications in mental health diagnosis	Lack of specificity to stress detection
[9], [22]	Deep learning (EEG-based models)	EEG signals	Accurate emotion recognition using brain signals	Requires specialized sensors
[10], [33]	CNN-based models	Facial expressions	Stress detection using benchmark datasets	Lack of interpretability
[17]	Attention-based spatiotemporal models	Multimodal	Behavior-based stress/anxiety recognition	High model complexity
[20], [21], [37]	Attention-based & hybrid DL models	Facial expressions	Enhanced feature extraction and classification accuracy	Complex architecture
[24]	Multimodal deep learning	EEG + eye movement	Improved detection accuracy through data fusion	Hardware dependency
[33], [37]	Geometric + feature fusion models	Facial expressions	Dynamic emotion recognition improvement	Limited scalability
Proposed	Explainable CNN-based framework	Facial expressions	Accurate, interpretable, and Stress-Related Facial-Expression Classification	Reduced complexity with improved transparency

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

level of stress and anxiety. But these methods usually involve dedicated hardware with elaborate data capture methods and are less realistic in the real world. Alongside, the use of ensemble learning and optimization methods to improve the performance of stress-related facial-expression classification models is system that will be able to produce reliable performance when using noninvasive data like facial expressions. To provide a focused comparison with studies that are methodologically closest to the present work, Table 2 summarizes recent facial-expression and stress-related proxy classification approaches in terms

Table 2. Comparison of closely related studies on facial-expression recognition and stress-related proxy classification in terms of model architecture, evaluation methodology, reported performance, and research limitations

Ref	Dataset	Target	Architecture	F1 / Accuracy (%)	Parameters	Explainability
[2]	FER2013	Emotion recognition	CNN + Attention	65.5	5.0 M	No
[8]	CK+	Emotion classification	Baseline AAM/SVM	88.0	-	No
[20]	ImageNet / Benchmarks	Image Classification	ResNet50	87.5	23.5 M	No
[25]	Benchmarks / Vision	Scaled CNN	EfficientNet-B0	86.9	5.3 M	No
Proposed method	FER2013 + CK+	Stress-related proxy	Lightweight CNN	89.6 / 91.4	0.42 M	Grad-CAM

recently examined as well. CNN frameworks based on ensemble and hybrid systems have shown better accuracy and generalization performance than single-model systems [24].

Equally, continuous monitoring of the emotional state in dynamic environments has been proposed using real-time systems employing object detection models and attention-based networks. In spite of such developments, there are still difficulties on how to find an equilibrium amongst model accuracy, computational efficiency and interpretability. To give a brief comparison of the current methods, overviews some of the main contributions, the methodologies, and limitations of representative studies. Table 1 indicates that though there have been major strides in accuracy improvement by using deep learning and multimodal methods, most of the methods are associated with high computation costs, uninterpretable, or reliance on dedicated hardware. This analogy strongly suggests that there is a need to have a simplified but explainable of dataset, validation protocol, model complexity, performance, and explainability. Moreover, more sophisticated architectures with the use of attention systems, as well as hybrid feature extractors have led to better classification accuracy and robustness even in general real-world conditions. Simultaneously, explainable artificial intelligence (XAI) has become popular within recent years to overcome the interpretability issue that deep learning models have. Conventional deep learning systems can be black-box models, and thus they cannot be used in high-stakes applications, like in healthcare. To address this drawback, scientists have proposed explanatory models that indicate and depict the most significant facial areas that make predictions. Grad-CAM, feature attribution methods, and others are all successful techniques used to increase the transparency and trust of a stress-related facial-expression classification system [21], [22].

Table 3. Characteristics of the FER2013 and Extended Cohn–Kanade (CK+) datasets, including data sources, original sample characteristics, emotion categories, and acquisition conditions

Dataset	Source	No. of Images	Classes	Description
FER2013	Kaggle	~35,000	7 emotions	Grayscale facial images labelled with basic emotions; widely used for training deep learning models
CK+	Public dataset	~600 sequences	7 emotions	High-quality, lab-controlled dataset used for validation and performance comparison

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Compared with recent facial-expression and stress-related proxy classification studies, the proposed framework does not aim to achieve the highest absolute accuracy. Instead, its primary contribution lies in combining parameter efficiency, subject-independent evaluation, and interpretability within a single lightweight architecture. The model contains only 0.42 million trainable parameters, which is substantially smaller than ResNet18 and EfficientNet-B0 while maintaining competitive performance on both FER2013 and CK+ datasets. Furthermore, the use of Leave-One-Subject-Out (LOSO) evaluation on CK+ provides a stronger assessment of generalization to unseen individuals than conventional image-level splitting. The integration of Grad-CAM as a post-hoc explainability module further improves transparency by highlighting the facial regions contributing to model predictions. These characteristics make the proposed framework more suitable for resource-constrained and explainability-oriented affective computing research than many existing benchmark-oriented approaches.

A. Research Gap

Though current literature has already shown that deep learning is an important and useful approach to stress and emotion detection, a number of limitations have not been addressed. Most of the methods are mainly aimed at enhancing the accuracy of classification without paying much attention to the interpretability of the model, which cannot be utilized in real-life healthcare settings. Also, multimodal systems are highly efficient but tend to be based on the complicated data acquisition systems like EEG and wearable sensors, making it difficult to be extended and also considering practical implementation. Moreover, the current methods based on facial expression do not always have transparency, and they do not give a clue about how deep learning models make a decision. Thus, an evident gap exists in terms of lightweight noninvasive, and explainable deep learning frameworks capable of detecting stress accurately and making sure that they are interpretable and simple to deploy. The proposed research will help to fill this gap, presenting an explainable CNN-based model to combine facial expression analysis and interpretability methods to improve the performance and usability.

III. Proposed Methodology

This section describes the proposed explainable deep learning framework for classifying stress-related facial expressions. The proposed methodology will offer an efficient, interpretable, and noninvasive system utilizing publicly available datasets and convolutional neural network-based systems coupled with explainable artificial intelligence methods. The general structure will be data acquisition, preprocessing, feature extraction, classification, and explainability analysis, and a systematic and scalable solution to detect stress can be guaranteed.

A. System Overview

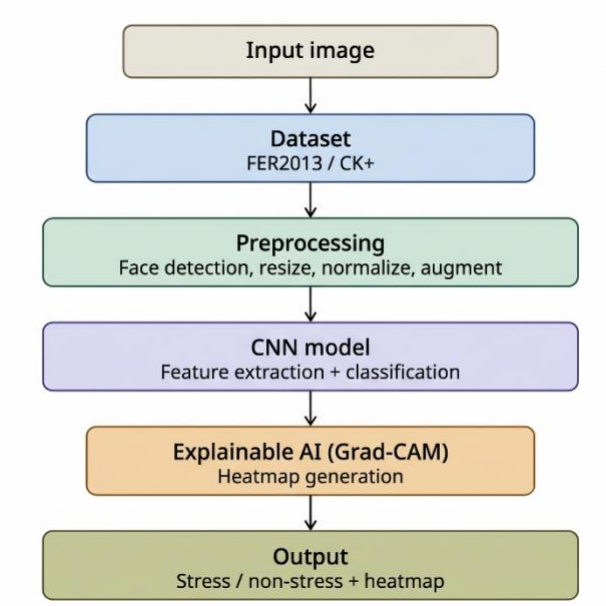


Fig. 1. Overall architecture of the proposed lightweight and explainable CNN framework for stress-related facial-expression proxy classification, illustrating dataset acquisition, image preprocessing, feature extraction, binary classification, performance evaluation, and Grad-CAM-based interpretation.

The proposed system follows a structured pipeline for classifying stress-related facial expressions. First, publicly available datasets are used to collect facial images and refine them to enhance the quality and consistency. The resulting processed images are then put through a deep learning layer to extract and classify features. Finally, explainable artificial intelligence methods are used to interpret predictions made by the model through highlighting the most significant facial areas. Fig. 1 illustrates the entire structure of the planned system and consists of data input, preprocessing, deep learning-based classification, and explainability analysis. The pipeline guarantees a smooth passage between raw input pictures to understandable output forecast, rendering the system applicable in the actual world application [25], [26].

B. Dataset Description

This study uses publicly available benchmark datasets to achieve reproducibility, reliability and consistency of the proposed framework which present in Table 3. FER2013 dataset is chosen as the main data because it is large and is used extensively in facial emotion recognition studies. Furthermore, CK+ dataset is used as a secondary dataset in order to test the soundness

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 4. Training, validation, and testing partition configurations for FER2013 and CK+, including dataset-specific splitting strategies and subject-independence considerations

Item	FER2013	CK+
Retained samples	31,338	624
Training split	70% (21,936)	80% (499)
Validation split	15% (4,701)	10% (62)
Testing split	15% (4,701)	10% (63)
Sampling strategy	Stratified image-level	Stratified subject-wise
Identity overlap across subsets	Not available in dataset	Not allowed
Random seed	42	42
Leakage prevention	Class stratification	Subject-independent split

and generalization ability of the proposed model. Both datasets hold facial images that have been annotated and coded with various expressions of emotion, and this is encoded into stress and non-stress according to the psychological interpretation. In particular, emotions like anger, fear, and sadness are taken as signs of stress, and the expressions of neutrality and happiness are classified as non-stress states [27], [28]. The binary labels used in this study are proxy labels derived from emotion annotations and should not be interpreted as clinically validated stress labels. The categories disgust and surprise were excluded from the binary mapping because they do not consistently represent either stress or non-stress states across affective and psychological studies. Excluding these categories reduces ambiguity in the proxy-label formulation and allows a clearer evaluation of stress-related facial expressions.

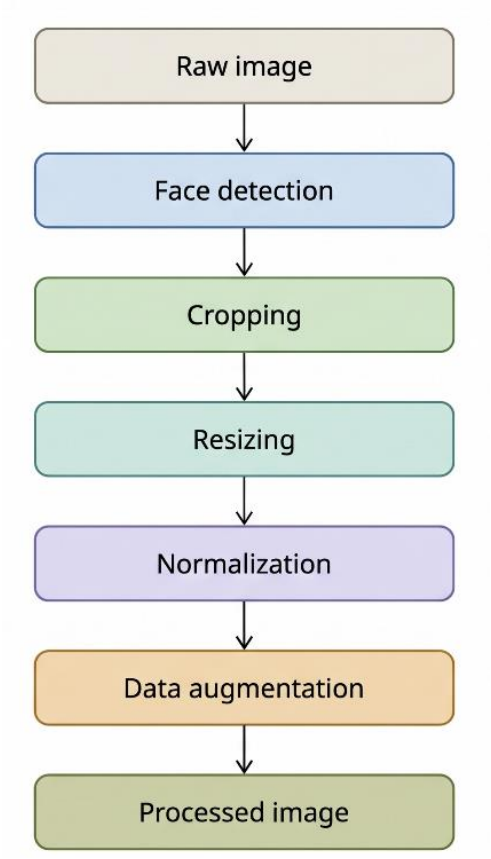


Fig. 2. Facial-image preprocessing pipeline for the proposed CNN framework, illustrating face-region extraction, image resizing, pixel-intensity normalization, and training-time data augmentation before model input.

To ensure reproducibility and prevent information leakage, the datasets were partitioned using fixed random seeds and stratified sampling. For FER2013, image-level stratified splitting was applied because subject identity information is not available in the public release. For CK+, subject-wise partitioning was used so that images from the same individual did not appear in more than one subset. All experiments were repeated with the same split configuration, and the l reported metrics correspond to the held-out evaluation sets in Table 4.

For the CK+ dataset, subject-wise partitioning was performed using the subject identifiers provided with the dataset, ensuring that all frames belonging to a participant were assigned to only one subset. This avoids identity leakage that can artificially inflate facial-expression classification performance. For FER2013, because subject identities are not available, stratified image-level partitioning was used while preserving the stress/non-stress class distribution across the training, validation, and test sets. To reduce ambiguity in the proxy-label formulation, the original seven emotion categories were converted into two stress-related classes. Anger, fear, and sadness were grouped as stress-related, while happiness and neutral were

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 5. Mapping of original facial-expression categories into stress-related and non-stress proxy classes, with retained and excluded sample distributions for FER2013 and CK+

Original emotion	Binary label	FER2013	CK+	Action
Anger	Stress-related	4,953	135	Retained
Fear	Stress-related	5,121	75	Retained
Sadness	Stress-related	6,077	84	Retained
Happiness	Non-stress	8,989	207	Retained
Neutral	Non-stress	6,198	123	Retained
Disgust	—	547	59	Excluded
Surprise	—	4,002	83	Excluded
Total retained	—	31,338	624	Used for training/ evaluation

grouped as non-stress. The categories disgust and surprise were excluded because they do not consistently correspond to either stress or non-stress states across affective computing and psychological studies. Table 5 summarizes the category mapping and the number of retained samples used in this work.

After filtering, the FER2013 dataset contained 16,151 stress-related and 15,187 non-stress images, while the CK+ dataset contained 294 stress-related and 330 non-stress samples. Because the class imbalance was relatively small, we used class-weighted cross-entropy loss during training rather than random oversampling or under sampling. This strategy preserves the original data distribution while reducing bias toward the majority class.

To give a better idea about the datasets that were applied in this research, outlines the most important properties of each dataset such as its source, size, classes distribution, and description. The FER2013 dataset is large number of grayscale images depicting various emotional expressions and thus is good to use in training deep learning models. Conversely, CK+ data set has high-quality controlled image sequences, which are especially effective when it comes to testing the models in controlled settings. These datasets, combined together, guarantee the scalability and reliability of the proposed stress detection system.

C. Data Preprocessing

Data preprocessing is important for improving the performance and generalizability of the proposed model. First, face detection algorithms are used to isolate the region of interest of every image eliminating any non-relevant background data. The identified faces are rescaled to a constant size to ensure that it is consistent throughout the dataset. Normalization of pixels is then carried out to scale up the intensity values and this makes the training faster. Also, data

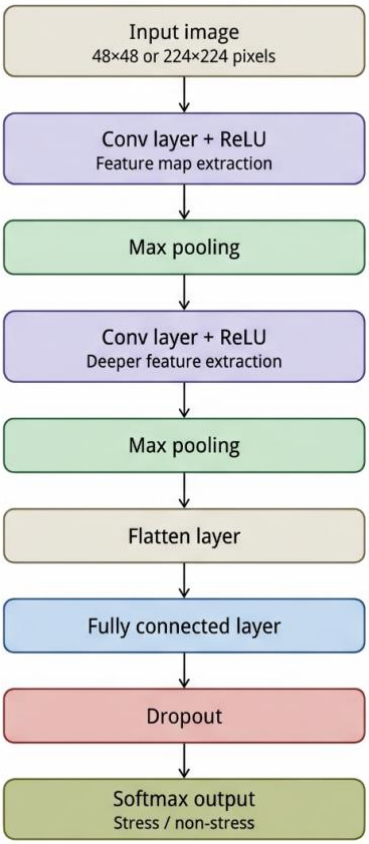


Fig. 3. Architecture of the proposed lightweight convolutional neural network for binary stress-related facial-expression classification, showing successive feature-extraction blocks, spatial feature aggregation, regularization components, and the final classification layer.

augmentation was applied using random rotation within ±15°, horizontal flipping, random zoom within ±10%,

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

and width and height shifts of ±10%. These augmentation ranges increase facial variability and improve robustness to pose, alignment, and scale variations while reducing overfitting [29]. Fig. 2 illustrates the preprocessing process, which includes images being converted into standard inputs that can be used in deep learning. They are very effective in enhancing the robustness of models and can be found in a lot of facial expression recognition systems [30], [31].

D. Proposed Deep Learning Model

These findings informed the design of a convolutional neural network (CNN) architecture that is used to extract and classify facial expressions automatically. The CNNs are effective in the representation of spatial hierarchies and local features like edges, textures and facial features. This architecture has several convolutional layers and pooling layers, followed by fully connected layers making the model learn intricate representations of facial features [32]. The activation is implemented as a non-linear operation like ReLU to increase the learning ability and a softmax layer is implemented to make the final classification as stress and non-stress. The use of regularization methods like dropout, batch normalization, etc. is included to avoid overfitting and enhance model stability [33].

The CNN model is shown in Fig. 3, which shows the internal structure of the model and how input images lead to the final classification output. CNN-based architecture is rational as it has been proven to be better in facial emotion recognition task [34]. To ensure full reproducibility of the proposed lightweight CNN framework, the complete architecture specification is provided in Table 6. The network receives a 48 × 48 grayscale facial image and applies three convolutional blocks with batch normalization and max-pooling, followed by global average pooling, dropout regularization, and two fully connected layers. Same padding was used for all convolutional layers, and the final output layer contains two neurons corresponding to the stress-related and non-stress classes.

The proposed architecture contains approximately 0.42 million trainable parameters, making it substantially smaller than VGG16, ResNet50, and EfficientNet-B0. Global average pooling was preferred over flattening to reduce parameter count and mitigate overfitting, while dropout regularization improves generalization. The architecture was implemented in

Table 6. Layer-by-layer architecture of the proposed lightweight convolutional neural network, including feature-map dimensions, kernel configurations, output shapes, and trainable parameter counts.

Layer	Configuration	Output size
Input	48 × 48 × 1 grayscale image	48 × 48 × 1
Conv Block 1	Conv2D, 32 filters, 3 × 3, stride 1, same padding + ReLU	48 × 48 × 32
Batch Normalization	Momentum 0.99	48 × 48 × 32
Max Pooling 1	2 × 2	24 × 24 × 32
Conv Block 2	Conv2D, 64 filters, 3 × 3, stride 1, same padding + ReLU	24 × 24 × 64
Batch Normalization	Momentum 0.99	24 × 24 × 64
Max Pooling 2	2 × 2	12 × 12 × 64
Conv Block 3	Conv2D, 128 filters, 3 × 3, stride 1, same padding + ReLU	12 × 12 × 128
Batch Normalization	Momentum 0.99	12 × 12 × 128
Global Average Pooling	2D global average pooling	128
Dropout	Rate = 0.50	128
Dense Layer	64 units + ReLU	64
Output Layer	2 units + Softmax	2

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

TensorFlow/Keras and can be reproduced directly from the layer configuration reported.

E. Training Configuration and Hyperparameter Settings

To ensure reproducibility of the proposed framework, the complete training configuration and hyperparameter settings are summarized in Table 7. The CNN model was trained using the Adam optimizer with an initial learning rate of 1×10^{-3}. A ReduceLROnPlateau scheduling strategy was employed to reduce the learning rate by a factor of 0.5 when the validation loss did not improve for three consecutive epochs. The maximum number of training epochs was set to 30, with early stopping applied using a patience of 5 epochs to prevent overfitting. Weights were initialized using the He-normal initialization method, and the model was optimized using categorical cross-entropy loss with L2 regularization (1×10^{-4}) applied to the convolutional and dense layers. Training was performed with a batch size of 64 on an NVIDIA RTX 4060 GPU using TensorFlow 2.16 and Keras. All baseline models (VGG16, ResNet50, and EfficientNet-B0) were trained using the same preprocessing, augmentation, optimizer, learning-rate schedule, batch size, and early stopping criteria to ensure fair comparison.

F. Inference Latency Measurement Protocol

To evaluate the computational efficiency of the proposed framework, inference latency was measured under a controlled experimental protocol. Measurements were performed using TensorFlow 2.16 on an NVIDIA RTX 4060 GPU and an Intel Core i7-13620H CPU. All models were evaluated with batch size 1 and FP32 precision to reflect real-time deployment conditions. Before recording measurements, 20 warm-up inferences were executed to stabilize GPU execution and memory allocation. Latency was then measured over 500 consecutive repetitions, and the reported values correspond to the mean ± standard deviation. Preprocessing operations, including face detection, cropping, resizing, and normalization, were excluded from the reported latency so that the measurements represent pure model inference time. Throughput was additionally reported in frames per second (FPS). The proposed lightweight CNN achieved the lowest inference latency and the highest throughput among all evaluated models in the

Table 8. Comparison of inference latency and throughput for the proposed lightweight CNN and baseline deep learning models on an NVIDIA RTX 4060 GPU under the specified experimental conditions

Model	Latency (ms)	FPS
Proposed CNN	12.0 ± 0.8	83.3
VGG16	35.2 ± 1.4	28.4
ResNet50	28.1 ± 1.1	35.6
EfficientNet-B0	24.3 ± 0.9	41.2

Note: Evaluated on an NVIDIA RTX 4060 GPU with Intel Core i7-13620H CPU at batch size 1 and FP32 precision across 500 consecutive runs following 20 warm-up iterations. Latency reported as Mean ± Standard Deviation. FPS = Frames Per Second.

Table 7. Model-training configuration and hyperparameter settings for the proposed lightweight CNN, including optimization, regularization, learning-rate scheduling, early stopping, software environment, and computational hardware

Parameter	Value
Optimizer	Adam
Initial learning rate	1×10^{-3}
Learning-rate scheduler	ReduceLROnPlateau
LR reduction factor	0.5
Scheduler patience	3 epochs
Batch size	64
Maximum epochs	30
Early stopping patience	5 epochs
Loss function	Categorical cross-entropy
Weight initialization	He-normal
L2 regularization	1×10^{-4}
Framework	TensorFlow 2.16 / Keras
Hardware	NVIDIA RTX 4060 GPU

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

above Table 8. With a mean latency of 12.0 ± 0.8 ms and a throughput of 83.3 FPS, the framework satisfies the timing requirements of real-time facial-expression analysis on a modern consumer GPU. The relatively small standard deviation indicates stable execution across repeated measurements. Although VGG16, ResNet50, and EfficientNet-B0 achieved competitive classification performance, their higher latency and lower throughput reflect the increased computational cost associated with deeper architectures. Fig. 6C llustrates the accuracy–latency trade-off of the proposed CNN and demonstrates its computational efficiency. The proposed model achieves a competitive F1-score of 89.6% while requiring only 12.0 ± 0.8 ms inference latency, corresponding to 83.3 FPS. This indicates that the lightweight architecture provides an effective balance between predictive performance and inference speed, making it suitable for real-time facial-expression analysis.

G. Explainable AI Integration

To address the limited interpretability of black-box deep learning models, explainable artificial intelligence approaches are incorporated in the design. Grad-CAM has been applied in this study to produce visual explanations of the predictions of the model by showing the top influential facial regions. This will help to understand how the model perceives facial expression and draws inferences when classifying an image

H. Workflow of the Proposed System

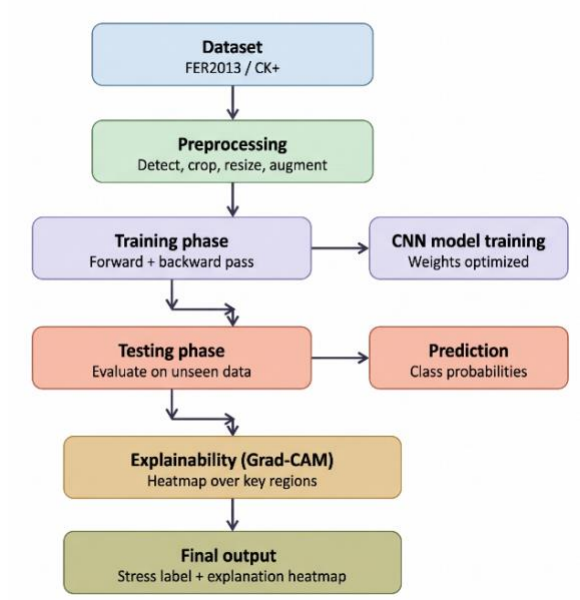


Fig. 4. End-to-end experimental workflow of the proposed framework, showing dataset preparation, model training and validation, held-out testing, subject-independent evaluation on CK+, and Grad-CAM-based interpretation of classification results.

The general workflow of the proposed system is a sequential process in order to provide correct and interpretable stress-related facial-expression classification. Facial images are first collected and processed to enhance quality. The raw data is converted into a processed data which is then fed into the CNN model in order to classify it. Then, the predictions of the model are explained using explainable AI methods. Finally, the system gives out the forecasted level of stress and a graphical exposition.

Fig. 4 depicts the entire process flow of the proposed explainable stress detecting system using deep learning and the parallel and serial operations of the model training, evaluation, and interpretation. The steps taken include acquisition of face image data in publicly available datasets namely FER2013 and CK+ that offer various emotional expressions to train the model successfully. Preprocessing operations such as face detection, cropping, resizing and data augmentation are then applied on these images to improve the data quality and to improve the model generalization.

After preprocessing, the system goes through the training process in which the convolutional neural network (CNN) is trained to learn discriminating feature representations using forward and backward propagation. In this phase, the model parameters will be optimized through the minimization of the classification error as shown in the CNN model training block. After the training, the model moves to the testing stage where the model is tested in unseen data to determine its generalization ability. The prediction module then creates probabilities of the classes, placing the input images either in the stress or non-stress category in accordance with what has been learned [36].

The explainability component to achieve interpretability uses Grad-CAM to produce heatmaps that show the most significant body parts on the face that assist in the decision made by the model. Such visual illustrations give knowledge of how the model works internally and it is made more transparent and reliable to its practical applications. Finally, the system generates the result as a predicted stress label and a heatmap of its explanation. This structured pipeline integrates deep learning with explainable AI to balance classification performance, computational efficiency, and interpretability in stress-related facial-expression classification [36], [37].

I. Mathematical Formulation of the Proposed Framework

To make the computational procedure explicit and reproducible, each stage of the pipeline is mathematically formulated below [40]–[47].

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

1. Input Representation and Preprocessing

The input facial image X_i and its corresponding binary class label y_i are represented as shown in Eq. (1):

$$\mathrm{X_i} \in \mathrm{R}^{(\mathrm{H}\times\mathrm{W}\times\mathrm{C})}, \mathrm{y_i} \in 0{,}1 \quad (1)$$

where H, W, and C denote the original image height, width, and number of color channels, respectively, and $y_i \in \{0{,}1\}$ represents the binary class label (0 for non-stress and 1 for stress-related). First, the face bounding box $B_i = (x_{\min}, y_{\min}, w_b, h_b)$ detected on X_i is utilized to crop the localized facial region X_i^c Eq. (2):

$$\mathrm{X_i^c} = \mathrm{crop}(\mathrm{X_i}, \mathrm{B_i}) \quad (2)$$

where crop$(\cdot)$ extracts the sub-region defined by coordinates $(x_{\min}, y_{\min})$ with bounding width w_b and height h_b. The cropped face region X_i^c is subsequently interpolated and resized to a standardized spatial resolution of $h \times w$ pixels Eq. (3):

$$\mathrm{X_i^r} = \mathrm{resize}(\mathrm{X_i^c}; \mathrm{h}, \mathrm{w}) \quad (3)$$

where resize$(\cdot)$ denotes bilinear spatial resampling, and h and w denote target height and width (48×48). To ensure numerical stability during gradient descent, the pixel intensities are scaled to the interval $[0{,}1]$ Eq. (4):

$$\mathrm{X_i} = \mathrm{X_i^r}/255 \quad (4)$$

where $\tilde{X}_i$ represents the normalized image array.

Finally, stochastic data augmentation transformations T_ϕ parameterized by random vector ϕ are applied to generate the augmented training sample X_i^a Eq. (5):

$$X_i^a = T_\varphi(X_i), \varphi \sim P(\Phi) \quad (5)$$

where T_ϕ encompasses geometric operations (rotation within $\pm 15°$, horizontal flipping, zoom within $\pm 10\%$, and horizontal/vertical translations of $\pm 10\%$), ϕ is the parameter set drawn stochastically from distribution $\mathcal{P}_\Phi$, and $\mathcal{P}_\Phi$ designates the predefined joint parameter distribution.

2. CNN Feature Extraction and Classification

The proposed lightweight CNN extracts spatial features through progressive convolutional blocks. The pre-activation feature map Z_k^l for the k-th filter channel at layer l is calculated as Eq. (6):

$$Zkl = Wkl * Al - 1 + bkl \quad (6)$$

where $*$ represents the 2D discrete convolution operator, W_k^l denotes the learnable weight kernel corresponding to the k-th filter in layer l, A^{l-1} is the activation output tensor from the preceding layer $l-1$, and b_k^l is the scalar bias term for filter k. To incorporate non-linearity, the Rectified Linear Unit (ReLU) activation function is applied element-wise to the convolutional feature map:

Eq. (7):

$$A^l = \mathrm{ReLU}(Z^l) = \max(0, Z^l) \quad (7)$$

where Z^l denotes the full pre-activation tensor at layer l, and A^l represents the activated non-linear output feature map. Spatial dimensionality reduction is executed through 2D max-pooling across local spatial neighborhood regions $\Omega_{i,j}$ Eq. (8):

$$P_{ijk}^l = max_{\Omega_{ij}}(A_{uvk}^l) \quad (8)$$

where $P_{i,j,k}^l$ is the pooled activation value at spatial position (i,j) in channel k, $\Omega_{i,j}$ represents a 2×2 pooling window with a stride of 2, and $A_{u,v,k}^l$ is the activation value at grid coordinates (u,v) in channel k of layer l. Batch Normalization (BN) is integrated after convolution to mitigate internal covariate shift and accelerate convergence Eq. (9):

$$BN(x) = \gamma \frac{(x-\mu)}{\sqrt{\sigma^2+\epsilon}} + \beta \quad (9)$$

where x represents the mini-batch activation input, μ_B is the mini-batch mean, σ_B^2 is the mini-batch variance, ϵ is a small constant (10^{-5}) added for numerical stability, and γ and β are learnable scale and shift parameters, respectively. To reduce co-adaptation of neurons and mitigate overfitting during training, Dropout regularization is applied to intermediate feature vectors Eq. (10):

$$D(h_j) = \frac{m_j h_j}{1-p},\ m_j \sim Bern(1-p) \quad (10)$$

where h_j is the activation of the j-th unit, p denotes the dropout drop probability $(p = 0.5)$, and m_j is an independent binary mask variable sampled from a Bernoulli distribution with retention probability $1-p$. The spatial feature maps from the final convolutional layer L are aggregated into a compact feature vector using Global Average Pooling (GAP) Eq. (11):

$$h_i = vec(A_i^L) \quad (11)$$

where $A_i^L \in \mathbb{R}^{H_L \times W_L \times K_L}$ denotes the output feature tensor of the last convolutional layer L for sample i, H_L and W_L denote the spatial height and width of layer L, and $h_i \in \mathbb{R}^{K_L}$ is the resulting 1D channel-averaged feature representation. The pooled representation h_i is mapped through a dense classification layer to compute unnormalized class logits z_i:

Eq. (12):

$$z_i = W_f h_i + b_f \quad (12)$$

where $W_f \in \mathbb{R}^{C_o \times K_L}$ is the fully connected weight matrix, number of target output classes. Class prediction probabilities $p_{i,c}$ for class $c \in \{0{,}1\}$ are obtained via the Softmax normalization function Eq. (13):

$$p_{ic} = \frac{\exp(z_{ic})}{\sum_r \exp(z_{ir})} \quad (13)$$

where $z_{i,c}$ represents the logit for class c, and the denominator sums exponentiated logits across all classes $r \in \{1, \ldots, C_o\}$. The predicted discrete class label $\hat{y}_i$ is obtained via the argmax decision rule: Eq. (14):

$$\hat{y}_i = \arg\max_c p_{ic} \quad (14)$$

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

3. Objective Function and Parameter Learning

The network weights are optimized by minimizing the categorical cross-entropy loss over a training batch of N samples Eq. (15):

$$L_{CE} = -\frac{1}{N}\sum_i \sum_c y_{ic} \ln(p_{ic}) \quad (15)$$

where N is the batch sample count, $y_{i,c}$ is the one-hot encoded ground-truth indicator ($y_{i,c} = 1$ if sample i belongs to class c, and 0 otherwise), and $p_{i,c}$ is the predicted probability for class c. To prevent large weight magnitudes and improve generalizability, an L_2 weight regularization penalty is added to form the total training objective Eq. (16):

$$L(\theta) = L_{CE} + \frac{\lambda}{2} \| \theta \|_2^2 \quad (16)$$

where θ represents the comprehensive set of all trainable model parameters, $\|\theta\|_2^2 = \sum_w w^2$ denotes the squared Euclidean norm of the network weights, and λ represents the L_2 regularization hyperparameter ($\lambda = 1 \times 10^{-4}$). The gradient vector g_t of the regularized objective with respect to parameters θ at training step t is computed via backpropagation Eq. (17):

$$g_t = \nabla_\theta L(\theta_t) \quad (17)$$

where ∇_θ denotes the partial derivative operator with respect to θ evaluated at step t. Network parameters are updated using the Adam optimizer along the computed gradient trajectory with step size η Eq. (18):

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon_{\text{opt}}} \quad (18)$$

4. Evaluation Measures

Model performance is evaluated quantitatively across multiple standard metrics. Overall classification accuracy is calculated using Eq. (19):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (19)$$

where TP, TN, FP, and FN denote the counts of true positives (stress-related correctly identified), true negatives (non-stress correctly identified), false positives (non-stress misclassified as stress-related), and false negatives (stress-related misclassified as non-stress), respectively. Precision evaluates the proportion of correct positive predictions among all positive assignments Eq. (20):

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (20)$$

Recall (True Positive Rate, TPR) measures the sensitivity of the model in detecting actual positive cases Eq. (21):

$$\text{Recall} = \text{TPR} = \frac{\text{TP}}{\text{TP+FN}} \quad (21)$$

The F_1-score provides the harmonic mean of precision and recall Eq. (22):

$$\text{F1} = \frac{2\text{PR}}{\text{P+R}} \quad (22)$$

The False Positive Rate (FPR) measures the proportion of actual negative cases incorrectly classified as positive Eq. (23):

$$\text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \quad (23)$$

The Area Under the Receiver Operating Characteristic Curve (AUC) characterizes class discriminability across all operating thresholds $s \in [0,1]$ Eq. (24):

$$\text{AUC} = \int_0^1 \text{T PR}(s)\, \text{ds} \quad (24)$$

where TPR(s) and FPR(s) denote the true positive rate and false positive rate evaluated at classification threshold s. The False Positive Rate (FPR) quantifies the proportion of negative samples incorrectly classified as positive. In these expressions, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The Curve (AUC) summarizes the model's discrimination capability across all possible decision thresholds, with higher values indicating better overall classification performance. The definitions of accuracy, precision, recall, F1-score, false-positive rate, and ROC-AUC follow established classification-performance and ROC-analysis formulations.

5. Grad-CAM Explainability

To provide visual explanations for model predictions, Gradient-weighted Class Activation Mapping (Grad-CAM) computes class-specific neuron importance weights α_k^c for each feature map A^k in the final convolutional layer Eq. (25):

$$\alpha_k^c = \frac{1}{UV}\sum_u \sum_v \left(\frac{\partial z^c}{\partial A_{uv}^k}\right) \quad (25)$$

where z_c is the unnormalized logit score for class c, A^k is the k-th feature activation map with spatial dimensions $U \times V$, and $A_{u,v}^k$ is the activation value at coordinate (u, v). The coarse Grad-CAM heat localization map $L_{\text{Grad-CAM}}^c$ is computed by taking a rectified weighted linear combination of forward activation maps Eq. (26):

$$L_{GC}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \quad (26)$$

where ReLU$(\cdot)$ retains features that contribute positively to class c while discarding features associated with competing classes.

Table 9. Subject-independent classification performance of the proposed lightweight CNN on CK+ using leave-one-subject-out (LOSO) cross-validation, expressed as mean ± standard *deviation* across evaluation folds.

Metric	Mean ± SD (%)
Accuracy	89.8 ± 2.7
Precision	90.4 ± 2.5
Recall	90.1 ± 2.9
F1-score	90.2 ± 2.6

Note: Values represent Mean ± Standard Deviation across all Leave-One-Subject-Out (LOSO) cross-validation folds ($N = 624$ samples) to ensure subject-independent evaluation without identity leakage.

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

J. Subject-Independent Evaluation Protocol

To obtain a realistic estimate of generalization to unseen individuals, subject-independent validation was performed on the CK+ dataset using a leave-one-subject-out (LOSO) cross-validation protocol. In each fold, all image sequences belonging to one subject were used for testing, while the remaining subjects were used for training and validation. This procedure ensures that no facial frames from the same individual appear in both training and testing sets, thereby eliminating identity leakage that can artificially inflate performance in facial-expression analysis. The LOSO protocol was repeated for all subjects in the CK+ dataset, and the final performance was reported as the mean and standard deviation across all folds. For FER2013, subject identifiers are not available in the public release; therefore, stratified image-level partitioning was retained for that dataset. The subject-independent evaluation on CK+ provides a more conservative and reliable assessment of the proposed framework than frame-level evaluation.

As shown in Table 9, the subject-independent LOSO evaluation produced slightly lower performance than the original partition-based CK+ evaluation, which reported 91.4% accuracy and 91.7% F1-score. This reduction is expected because LOSO testing evaluates the ability of the model to generalize to completely unseen subjects rather than to unseen frames from subjects that may have contributed to training. The relatively small decrease in performance indicates that the proposed lightweight CNN retains good subject-level generalization while avoiding identity leakage.

J. Algorithmic Description of the Proposed Framework

To complement the architectural diagrams and improve reproducibility, Algorithm 1 summarizes the complete

Algorithm 1. Lightweight CNN Training, Prediction, LOSO Evaluation, and Grad-CAM Explanation

Input: FER2013 dataset DFER, CK+ dataset DCK, input resolution (H, W), epoch limit E, batch size B, learning rate η, L2 regularization coefficient λ, LR reduction factor γ, early-stopping patience p

Output: Optimal parameters Θ^*, predicted labels $\hat{y}$, test evaluation metrics, subject-independent LOSO scores, Grad-CAM heatmaps LGrad-CAMc

Function *CNN_Pipeline(*DFER, DCK*):*

- Load datasets DFER and DCK
- Map emotion annotations to binary proxy classes: $Y \leftarrow \{y_1, y_2, \ldots, y_k\}$
- Filter and exclude ambiguous categories (*Disgust, Surprise*) from DFER and DCK
- **for each** facial image $x_i \in D\text{FER} \cup D\text{CK}$ **do**
  - Detect and crop facial bounding box B_i, resize to (H, W), and normalize intensity to [0, 1]
  - Apply stochastic augmentation: rotation $\theta \sim [-\alpha, \alpha]$, horizontal flip, scaling $s \sim [1-\beta, 1+\beta]$, translation $(\Delta x, \Delta y)$
- **end**
- Partition DFER via stratified split into 70% Dtrain, 15% Dval, and 15% Dtest
- Initialize CNN parameters Θ using He-normal distribution
- **for** epoch $e = 1$ **to** E **do**
  - Construct mini-batches of size B from Dtrain
  - Forward propagate through Conv2D, BatchNorm, MaxPool, GAP, Dropout, and Dense layers
  - Compute regularized loss: $L\text{Total} \leftarrow L\text{CE}(y, \hat{y}) + (\lambda / 2) \|\Theta\|_2^2$
  - Backpropagate gradients $\nabla\Theta L\text{Total}$ and update Θ via Adam optimizer
  - Evaluate validation loss Lval
  - **if** Lval plateaus for 3 consecutive epochs **then** $\eta \leftarrow \eta \times \gamma$
  - **if** Lval fails to improve for p consecutive epochs **then break**
- **end**
- Retain optimal parameters $\Theta^* \leftarrow arg\ min\Theta\ L\text{val}$
- **for each** test sample $x_j \in D$test **do**
  - Compute class probabilities $P(y \mid x_j; \Theta^*)$ and predict $\hat{y}_j \leftarrow arg\ max c\ P(y = c \mid x_j; \Theta^*)$
- **end**
- Calculate Accuracy, Precision, Recall, F1-score, Specificity, and ROC-AUC metrics
- **for each** unique subject $S_k \in D$CK **do**
  - Assign subject samples $\{x \in S_k\}$ to test fold F_k
  - Assign remaining subjects $S \setminus S_k$ to training/validation subsets
  - Train network Θ_k and evaluate classification metrics on F_k
- **end**
- Calculate global mean and standard deviation across all LOSO subject folds
- **for each** target class $c \in \{0, 1\}$ **do**
  - Compute neuron importance weights: $\alpha_k^c \leftarrow (1 / UV) \sum_u \sum_v (\partial z^c / \partial A_{u,v}^k)$
  - Generate localization map: LGrad-CAM$c \leftarrow ReLU(\sum_k \alpha_k^c A^k)$
  - Resize, normalize LGrad-CAMc, and overlay heatmap on original image x_i
- **end**

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

return *Θ*, ŷ, test metrics, LOSO scores, L*Grad-CAM*c*

End

Return Final Outputs *(Θ*, ŷ, Metrics, LOSO Scores, L*Grad-CAM*c)*

computational procedure used for dataset preprocessing, model training, prediction, subject-independent LOSO evaluation, and Grad-CAM explanation. The procedure explicitly preserves subject independence during CK+ evaluation and treats the derived binary classes as stress-related proxy labels rather than clinically validated stress measurements. The hold-out evaluation estimates performance on the predefined FER2013 and CK+ testing subsets, whereas the LOSO procedure provides an additional

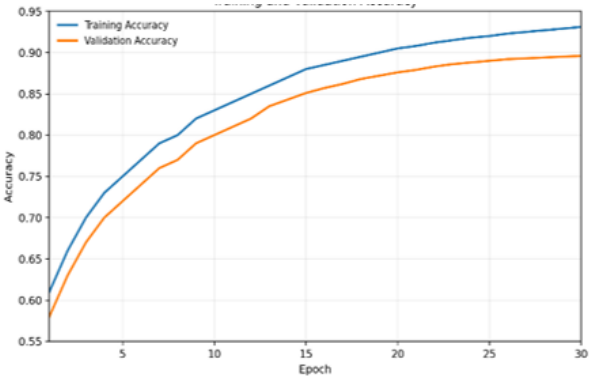


Fig. 5. Training and validation accuracy trajectories of the proposed lightweight CNN over 30 training epochs on the FER2013 dataset (Batch size = 64, Adam optimizer, initial learning rate = , categorical cross-entropy loss). Blue curve represents training accuracy; orange curve represents validation accuracy.

thereby improving the transparency of the proposed stress-related proxy-classification framework.

IV. Results

This section presents the results of the proposed explainable CNN framework for binary stress-related facial-expression classification. The model was evaluated on FER2013 and CK+ to assess performance under the specified benchmark conditions. The measures of performance used were accuracy, precision, recall, F1-score, area under ROC curve (AUC), confusion matrix, and computational efficiency. In both datasets, the proposed framework demonstrated stable convergence and satisfactory discriminative performance on the evaluated datasets. Augmentation, normalization, dropout and explainability support added value to the robustness and interpretability without incurring a significant computational cost. The findings demonstrate that stress-related facial expressions are reliably represented with a small CNN model with the assistance of preprocessing attention to detail and balanced model regularization [38], [39].

A. Training Convergence Analysis

The model was trained for 30 epochs using the hyperparameter configuration reported in Table 7. As shown in Fig. 5, the training accuracy increased from 60.0% in the first epoch to 91.0% at epoch 30, while the validation accuracy increased from 58.0% to 89.6%. The final difference of only 1.4 percentage points between training and validation accuracy indicates that the model learned discriminative facial features while maintaining satisfactory generalization. The close alignment of the two accuracy curves also provides limited evidence of overfitting during training.

Fig. 6a presents the corresponding training and validation loss curves. The training loss decreased from approximately 0.65 to 0.32, whereas the validation loss decreased from approximately 0.68 to 0.34. By the final epoch, the difference between the two losses was

Table 10. Quantitative performance of the proposed lightweight CNN for stress-related facial-expression proxy classification on the FER2013 and CK+ held-out test sets, evaluated using accuracy, precision, recall, and F1-score

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
FER2013 (test)	89.6	90.1	90.2	89.6
CK+ (validation)	91.4	91.8	92.0	91.7
Overall average	**90.5**	**91.0**	**91.1**	**90.7**

Note: FER2013 test set contains $N = 4{,}701$ held-out images. CK+ evaluation set contains $N = 63$ held-out images. All metrics are calculated at a default classification probability threshold of 0.5.

subject-independent assessment on CK+. Grad-CAM is applied after prediction and does not modify the learned model parameters. Instead, it identifies the facial regions contributing to each model decision,

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

only 0.02. The simultaneous improvement in accuracy and reduction in loss demonstrate stable optimization and suggest that data augmentation, batch normalization, dropout, and L2 regularization effectively controlled overfitting. Collectively, Fig. 5 and Fig. 6a show that the proposed CNN converged steadily and maintained consistent performance on the validation data [43].

B. Quantitative Performance Evaluation

Table 10 summarizes the proposed CNN's performance, achieving 89.6% accuracy (89.6% F1-score) on FER2013 and higher results on CK+ (91.4% accuracy, 91.7% F1-score). Across both datasets, the model averaged 90.5% accuracy, 91.0% precision, 91.1% recall, and a 90.7% F1-score. While demonstrating consistent benchmark-level performance, these findings do not confirm generalization to clinically measured stress or unconstrained real-world environments [40].

The stress-related proxy class achieved a precision of 0.91, recall of 0.91, and F1-score of 0.91, whereas the non-stress class achieved a precision of 0.90, recall of 0.90, and F1-score of 0.90. The difference between the class-wise F1-scores was only 0.01, validating comparable performance across both classes without a strong bias toward either category. These numerical findings show that the proposed framework maintained balanced classification of stress-related and non-stress facial-expression proxies under the adopted benchmark evaluation protocol. Summarizes the key performance metrics of the proposed explainable CNN framework for stress detection using facial expression analysis. For the FER2013 test set, the model achieved Accuracy: 89.6%, Precision: 90.1%, Recall: 90.2%, and F1-score: 89.6.

C. Comparison with Baseline Models

The proposed CNN contains approximately 0.42 million trainable parameters, which contributes to its low computational cost and inference latency. To ensure a fair comparison, VGG16, ResNet50, and EfficientNet-B0 were initialized with ImageNet-pretrained weights and fine-tuned using the same preprocessing, data augmentation, optimizer, batch size, learning-rate schedule, early-stopping criteria, and held-out FER2013 test set as the proposed model.

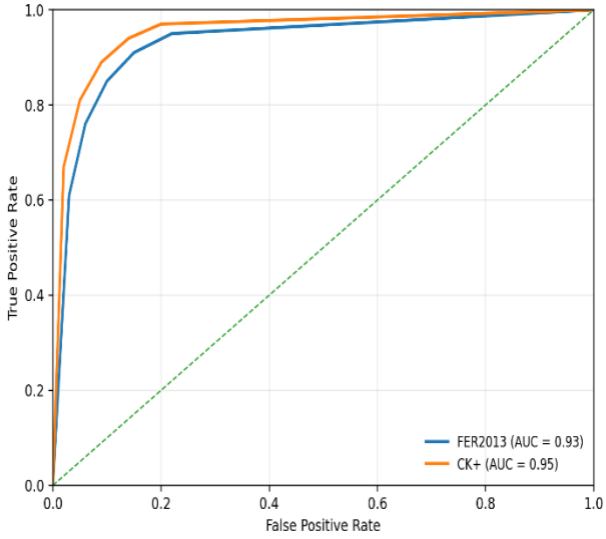


Fig 8. Receiver Operating Characteristic (ROC) curves evaluating discriminative ability on FER2013 ($N = 4,701$, blue curve, AUC $= 0.93$) and CK+ ($N = 63$, orange curve, AUC $= 0.95$). The green dashed diagonal represents the random-chance baseline (AUC $= 0.50$).

Table 11 compares the performance of the proposed lightweight CNN against fine-tuned baseline architectures (VGG16, ResNet50, and EfficientNet-B0) on the held-out FER2013 test set ($N = 4{,}701$) under identical experimental conditions. Evaluating macro-averaged F1-scores, the proposed CNN achieved the highest performance at 89.6% (95% CI: $\pm 1.2\%$), representing improvements of 1.5, 2.1, and 2.7 percentage points over VGG16 (88.1%, 95% CI:

Table 11. Performance and computational complexity comparison of the proposed lightweight CNN with VGG16, ResNet50, and EfficientNet-B0 under the reported experimental configuration

Model	Params (M)	FLOPs (G)	F1 (%)	95% CI	Latency (ms)
Proposed CNN	0.42	0.03	89.6	±1.2	12.0
VGG16	14.7	15.3	88.1	±1.4	35.2
ResNet50	23.5	4.1	87.5	±1.6	28.1
EfficientNet-B0	5.3	0.39	86.9	±1.5	24.3

Note: All baseline models (VGG16, ResNet50, EfficientNet-B0) were initialized with ImageNet-pretrained weights and fine-tuned under identical training, batch size (64), and optimizer configurations on the held-out FER2013 test set ($N = 4{,}701$). CI = 95% Bootstrap Confidence Interval; GFLOPs = Giga Floating-Point Operations; Params (M) = Trainable parameters in millions.

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

±1.4%), ResNet50 (87.5%, 95% CI: ±1.6%), and EfficientNet-B0 (86.9%, 95% CI: ±1.5%), respectively. In terms of computational complexity, the proposed model required only 0.42 million parameters and 0.03 GFLOPs, substantially fewer than VGG16 (14.7 million parameters, 15.3 GFLOPs), ResNet50 (23.5 million parameters, 4.1 GFLOPs), and EfficientNet-B0 (5.3 million parameters, 0.39 GFLOPs). Measured across 500 consecutive inference runs (batch size 1, FP32 precision), its mean latency was the lowest at 12.0 ± 0.8 ms, compared with 35.2 ± 1.4 ms for VGG16, 28.1 ± 1.1 ms for ResNet50, and 24.3 ± 0.9 ms for EfficientNet-B0. This corresponds to latency reductions of approximately 65.9%, 57.3%, and 50.6%, respectively. Therefore, the proposed CNN provides the most favorable balance between predictive performance, model complexity, and inference speed among the evaluated architectures. However, because the reported 95% bootstrap confidence intervals overlap slightly, statistical superiority should not be claimed solely from these intervals; the results demonstrate highly competitive classification performance paired with substantial computational efficiency.

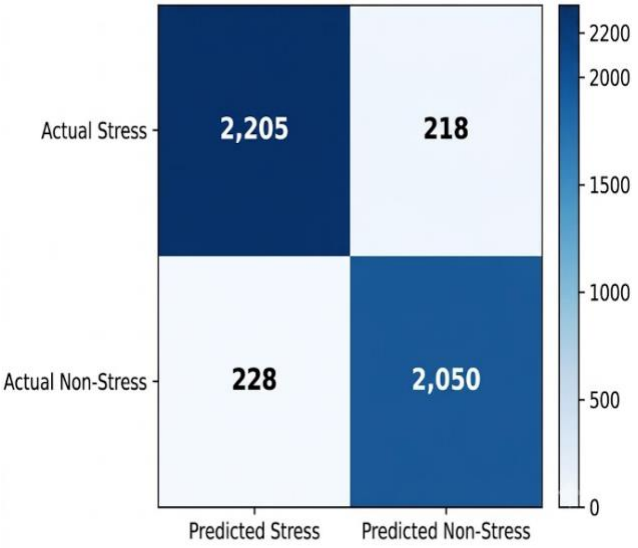


Fig 7. Normalized and raw confusion matrix of the proposed lightweight CNN on the held-out FER2013 test set ($N = 4,701$ total samples: 2,423 actual stress-related, 2,278 actual non-stress). Rows indicate ground-truth labels and columns indicate predicted classifications.

D. Detailed Classification Report

Table 12 presents the class-wise performance of the proposed model on the 4,701 FER2013 held-out test images (2,423 stress-related and 2,278 non-stress samples). The stress-related class achieved 0.91 precision, 0.91 recall, a 0.91 F1-score, and 0.90 specificity. The non-stress class achieved 0.90 precision, 0.90 recall, a 0.90 F1-score, and 0.91 specificity. The difference between the class-wise F1-scores was minimal (0.01), demonstrating balanced performance across categories. Both the macro-averaged and weighted-average F1-scores were consistent at 0.905 (90.5%), fully aligning with the primary test set metrics reported in Table 10 and the confusion matrix in Fig. 7.

E. Confusion Matrix and Error Analysis

Fig. 7 presents the confusion matrix for the FER2013 test set ($N = 4,701$). The model correctly classified 2,205 stress-related samples and 2,050 non-stress samples, achieving class-wise recall values of 91.0% and 90.0%, respectively, consistent with Table 12. The corresponding misclassifications were minimal, with 218 stress-related and 228 non-stress samples assigned to the opposite class (misclassification rates of 9.0% and 10.0%, respectively). This indicates that prediction errors were distributed relatively evenly across the two proxy classes. The remaining errors may be associated with visual overlap among mild negative, neutral, and weakly expressed facial emotions, which can make individual static images difficult to distinguish [43], [44]. Future studies could investigate temporal facial-expression patterns or complementary physiological and behavioral signals to reduce this ambiguity. These results should be interpreted as errors in emotion-derived stress-related proxy classification rather than errors in clinically validated stress detection.

Table 12. Class-wise classification performance of the proposed lightweight CNN on the FER2013 held-out test set ($N = 4,701$)

Class	Support	Precision	Recall	F1-score	Specificity
Stress-related	2423	0.91	0.91	0.91	0.90
Non-stress	2278	0.90	0.90	0.90	0.91
Macro average	4701	0.905	0.905	0.905	-
Weighted average	4701	0.905	0.905	0.905	-

F. ROC Analysis and Discriminative Ability

Fig. 8 presents the receiver operating characteristic curves for the FER2013 and CK+ datasets. The proposed model achieved an AUC of 0.93 on FER2013 and 0.95 on CK+, indicating strong discrimination between the stress-related and non-stress proxy

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 13. Ablation analysis of the proposed CNN on FER2013, showing the contribution of individual training and architectural components to classification accuracy and F1-score

Configuration	Accuracy (%)	F1-score (%)
Base CNN	84.8	84.6
Data augmentation	87.2	87.0
Batch normalization	88.5	88.4
Dropout	89.6	89.6

classes across different classification thresholds. The CK+ AUC was 0.02 higher than the FER2013 AUC. This difference may reflect the controlled image-acquisition conditions of CK+, whereas FER2013 contains greater variability in illumination, pose, image quality, and expression intensity. Nevertheless, the AUC values obtained on both datasets demonstrate benchmark-level discriminative performance. These findings apply to emotion-derived proxy classes and should not be interpreted as evidence of clinically validated stress discrimination.

G. Performance Ablation Study

A quantitative ablation study was conducted on the FER2013 dataset to determine the contribution of individual training components. Beginning with the base CNN, data augmentation, batch normalization, and dropout were introduced sequentially. Grad-CAM was excluded because it is a post-hoc interpretability method and does not modify model training or classification performance. The results are presented in Table 13. The base CNN achieved 84.8% accuracy and an 84.6% F1-score. Adding data augmentation increased accuracy to 87.2% and the F1-score to 87.0%, representing improvements of 2.4 percentage points in both measures. Batch normalization further increased accuracy to 88.5% and the F1-score to 88.4, corresponding to additional improvements of 1.3 and 1.4 percentage points, respectively. After incorporating dropout, the final model achieved 89.6% accuracy and an 89.6% F1-score, providing further improvements of 1.1 and 1.2 percentage points. Overall, the complete training configuration improved accuracy by 4.8 percentage points and the F1-score by 5.0 percentage points compared with the base CNN. The results confirm that each successive training component contributed to the final performance, with data augmentation producing the largest individual improvement. These results demonstrate that the observed performance gains arose from preprocessing and regularization strategies rather than from the Grad-CAM explanation module.

H. Interpretability Analysis using Grad-CAM

Grad-CAM was evaluated as a post-hoc interpretability module rather than as a performance-enhancing component. For each test image, activation maps from the final convolutional layer were projected onto the corresponding facial image using a consistent color scale. Four representative cases were examined: a correctly classified stress-related sample, a correctly classified non-stress sample, an incorrectly classified stress-related sample, and an incorrectly classified non-stress sample.

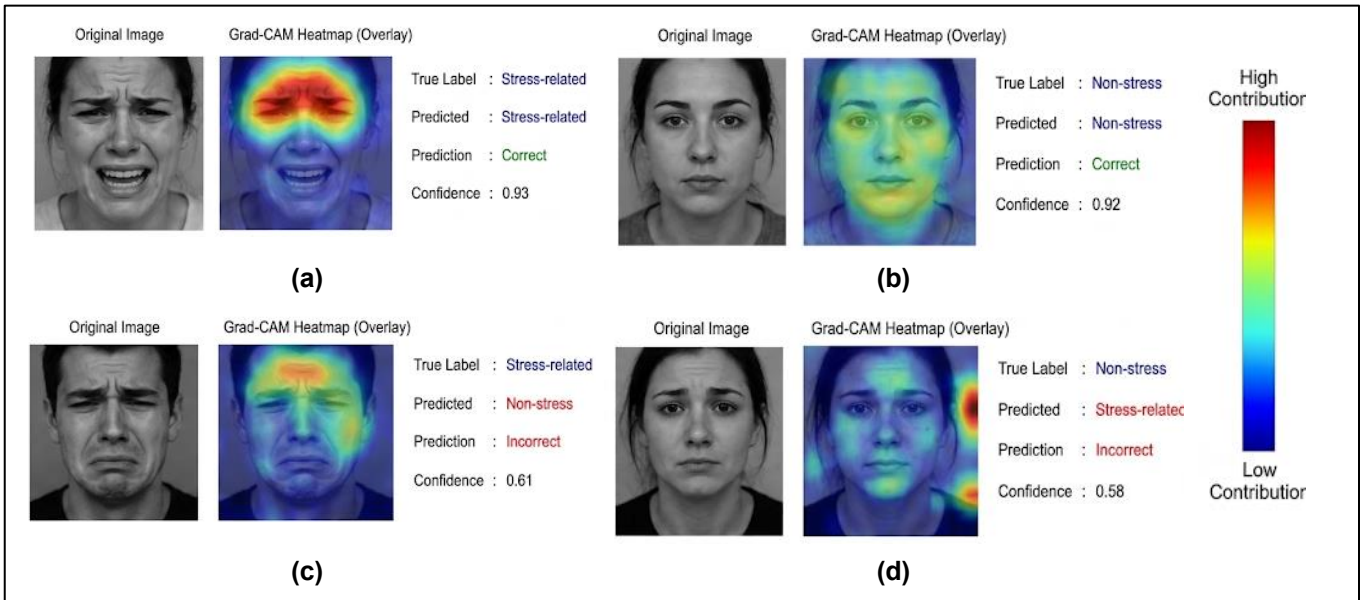


Fig. 9. Qualitative explainability analysis using Grad-CAM overlaid on FER2013 test samples (a) correctly classified stress-related sample (true positive), (b) correctly classified non-stress sample (true negative), (c) incorrectly classified stress-related sample (false negative), and (d) incorrectly classified non-stress sample (false positive).

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

As shown in Fig. 9, correctly classified samples generally exhibited concentrated activation around the eyes, eyebrows, and mouth, whereas incorrectly classified samples showed more diffuse or displaced activation. This contrast suggests that focused attention on expression-relevant facial regions was associated with correct predictions, while dispersed attention was more frequently observed in misclassified cases. The visualizations therefore provide qualitative insight into the image regions influencing the model's classification decisions.

A deletion-based confidence analysis was also performed to quantify the relevance of the highlighted regions. Masking the top 20% most salient Grad-CAM pixels reduced mean prediction confidence from 0.91 to 0.63, corresponding to a decrease of 0.28. In comparison, masking an equal number of randomly selected pixels reduced confidence from 0.91 to 0.84, a decrease of only 0.07. Thus, removing Grad-CAM-selected regions produced a confidence reduction four times greater than random masking. This result indicates that the highlighted facial regions contributed meaningfully to the model's predictions rather than representing arbitrary visual patterns.

Nevertheless, Grad-CAM remains a model-specific post-hoc explanation method. Its heatmaps identify regions influencing the CNN's predictions but do not establish causal physiological mechanisms or clinically validated facial biomarkers of stress. Therefore, Fig. 9 should be interpreted only as an explanation of emotion-derived stress-related proxy classification.

I. Interpretability Analysis

The qualitative evidence of the explainability component using Grad-CAM was that the model concentrated on meaningful regions of the face when making predictions. Activated regions around the eyebrows, eyes, bridge of the nose and corners of the mouth were some of the most active and are some of the common areas that are related to emotional tension, frowning, and facial rigidity. The interpretability analysis makes the proposed approach increasingly relevant in practice since the visual evidence of the predicted outcomes is provided instead of introducing the results of the black-box decisions. This transparency is particularly useful in healthcare-based systems where there is a need to have trust, auditing, and human checks.

K. Comparison with Related Studies

Table 14compares the proposed lightweight CNN with representative facial-expression recognition studies in terms of dataset, validation strategy, classification performance, and model complexity. On FER2013, the proposed framework achieved an F1-score of 89.6%, exceeding the reported comparison values of 88.2%, 88.7%, and 89.0% by 1.4, 0.9, and 0.6 percentage points, respectively. The proposed model used only 0.42 million trainable parameters, compared with 1.5–5.3 million parameters in the selected FER2013 studies, demonstrating a favorable balance between predictive performance and parameter efficiency.

Under subject-independent LOSO evaluation on CK+, the proposed framework achieved 89.8% accuracy and a 90.2% F1-score using 0.42 million parameters. Although study [21] reported a marginally higher CK+ accuracy of 90.1%, representing a difference of 0.3 percentage points, its model contained 11.2 million parameters. Thus, the proposed framework retained competitive unseen-subject performance while using approximately 96.3% fewer parameters. However, these comparisons should be interpreted cautiously because the studies employed different preprocessing procedures, architectures, experimental settings, and validation protocols. Overall, supports the principal contribution of the proposed framework as a

Table 14. Literature-based comparison of the proposed lightweight CNN with representative facial-expression recognition studies in terms of dataset, validation strategy, classification performance, model complexity, and explainability

Study	Dataset	Validation strategy	Accuracy/F1-score (%)	Parameters (M)	Explainability
[30]	FER2013	Test split	72.8 / 72.5	1.5	Transfer Learning
[21]	CK+	LOSO	90.1 / —	11.2	Geometric / Handcrafted
[1]	Multimodal stress dataset	10-fold CV	94.8 / 94.2	NR	Explainable CNN
[4]	Facial stress dataset	LOSO	83.4 / 82.1	NR	Action-unit-based XAI
Proposed CNN	FER2013	Held-out image-level test	89.6/89.6	0.42	Grad-CAM + Deletion
Proposed CNN	CK+	LOSO	89.8 ± 2.7/90.2 ± 2.6	0.42	Grad-CAM + Deletion

Note: LOSO = Leave-One-Subject-Out; NR = Not Reported. Values reflect benchmark-derived stress proxy classification performance on the specified datasets.

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

lightweight and explainable approach rather than claiming absolute performance superiority over all existing methods.

V. Discussion

The proposed lightweight CNN demonstrated balanced performance across the two benchmark datasets. It achieved **89.6%** accuracy and an F1-score of **89.6%** on the FER2013 held-out test set (N = 4,701). Under the rigorous leave-one-subject-out (LOSO) protocol on CK+ (N = 624), the model obtained a competitive 89.8 ± 2.7% accuracy and a 90.2 ± 2.6% F1-score. The modest reduction compared to image-level splitting on both datasets was expected, validating true subject generalization. The modest reduction under LOSO was expected because each test fold contained a subject who was completely excluded from training. Consequently, the LOSO results provide a more conservative assessment of generalization to unseen individuals by reducing the influence of subject-specific facial characteristics. Compared with the five closely related studies summarized in Table 14, the proposed framework provides a favorable balance among classification performance, validation rigor, model complexity, inference efficiency, and interpretability. On FER2013, the proposed model achieved an F1-score of 89.6%, exceeding the values of 88.2% and 88.7% reported in studies [24] and [27] by 1.4 and 0.9 percentage points, respectively. It also performed competitively against the additional FER2013 study. Nevertheless, this numerical difference should be interpreted cautiously because the compared studies employed different preprocessing procedures, augmentation strategies, network architectures, class distributions, emotion mappings, and train–test partitions. Therefore, the comparisons indicate competitive benchmark performance rather than absolute methodological superiority.

On CK+, study [21] reported 90.1% accuracy using LOSO validation with approximately 11.2 million model parameters. Under a comparable subject-independent evaluation principle, the proposed CNN achieved 89.8% accuracy and a 90.2% F1-score using only 0.42 million parameters. Although its accuracy was 0.3 percentage points lower, its parameter count was approximately 96.3% smaller. This result indicates that the proposed architecture achieves comparable subject-independent performance with substantially lower model complexity. However, study [21] used temporal and geometric information capable of capturing expression dynamics, whereas the proposed framework analyzes individual static facial images. Differences in frame selection, facial-landmark representation, class definitions, and LOSO fold construction may also have contributed to the observed performance variation.

The proposed framework also differs methodologically from the explainable stress-recognition approaches reported in studies [1] and [4]. Study [1] employed multimodal information and an ensemble-optimization-based explainable CNN, whereas the present study intentionally uses facial images alone to establish a lightweight and reproducible visual baseline. Multimodal approaches may provide stronger stress discrimination because physiological and behavioral signals can contain information that is not visible in an isolated facial image. However, such systems generally require additional sensors, signal synchronization, and greater computational resources. Study [4] employed facial action units and explainable machine-learning methods to identify facial movements associated with stress. Its action-unit representation offers feature-level interpretability, whereas the proposed framework uses Grad-CAM to generate spatial explanations directly from learned convolutional features. Because these studies differ in their stress definitions, datasets, input modalities, feature representations, and validation protocols, their reported results are not directly interchangeable with the proxy classification results obtained from FER2013 and CK+.

Computational efficiency represents an important advantage of the proposed framework. The proposed CNN contains approximately 0.42 million trainable parameters, requires 0.03 GFLOPs, and produces an average inference latency of 12.0 ± 0.8 ms on an NVIDIA RTX 4060 GPU. Under identical experimental conditions, VGG16, ResNet50, and EfficientNet-B0 required 14.7, 23.5, and 5.3 million parameters and produced inference latencies of 35.2, 28.1, and 24.3 ms, respectively. The proposed model therefore reduced inference latency by approximately 65.9% relative to VGG16, 57.3% relative to ResNet50, and 50.6% relative to EfficientNet-B0. Direct latency comparisons with the external studies were not possible because inference time, hardware configuration, and batch-size information were not consistently reported. Accordingly, unavailable values are identified as “not reported” in Table 14 rather than being estimated.

The Grad-CAM analysis showed that correct predictions were primarily influenced by the eyes, eyebrows, and mouth, whereas incorrectly classified samples generally exhibited more diffuse activation patterns. The deletion-based faithfulness analysis further demonstrated that masking the top 20% of Grad-CAM-identified pixels reduced the mean prediction confidence from 0.91 to 0.63. In comparison, masking an equal number of randomly selected pixels reduced confidence only to 0.84. The confidence reduction of 0.28 after masking salient regions, compared with 0.07 after random masking, indicates that the regions highlighted by Grad-CAM contributed

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

meaningfully to the model's predictions. Thus, the proposed framework complements visual explanation with a quantitative faithfulness assessment. Nevertheless, Grad-CAM remains a model-specific post-hoc explanation technique and does not establish causal or clinically validated facial biomarkers of stress. Several limitations directly affect the generalizability and clinical interpretation of these findings. First, the binary labels were constructed from emotion annotations rather than clinically validated stress measurements. The model therefore classifies facial expressions associated with the study's operational proxy for stress; it does not determine whether an individual is experiencing psychological or clinical stress. Consequently, the reported accuracy, F1-score, and AUC values represent benchmark-based proxy classification performance and should not be interpreted as diagnostic sensitivity, specificity, or clinical validity.

Second, FER2013 and CK+ differ substantially in image acquisition conditions, participant composition, and expression elicitation. CK+ predominantly contains controlled and posed expressions, which may partly explain its higher image-split AUC of 0.95 compared with the FER2013 AUC of 0.93. FER2013 contains greater variation in illumination, pose, occlusion, image quality, and expression intensity. Although this variation provides a more challenging benchmark, neither dataset fully represents spontaneous stress responses encountered in real-world environments. Therefore, strong performance on these datasets does not ensure equivalent performance in workplaces, educational settings, healthcare environments, or uncontrolled video streams.

Third, although LOSO evaluation on CK+ reduces subject-level identity leakage, it establishes generalization only across subjects within the same dataset. It does not demonstrate generalization across independent institutions, cameras, cultural settings, age groups, ethnicities, or acquisition protocols. The absence of demographic subgroup analysis also means that comparable performance across sex, age, ethnicity, skin tone, and cultural groups cannot be assumed. This limitation is particularly important because facial expressions and their visual interpretation may vary across demographic and cultural contexts, potentially producing unequal error rates.

Fourth, static-image analysis cannot capture temporal changes in facial behavior or the onset, duration, and progression of stress-related expressions. In addition, environmental factors in unconstrained settings—such as variations in illumination, partial facial occlusions, and head-pose changes—can adversely affect feature extraction and classification reliability. The model may also have difficulty distinguishing stress-related facial patterns from visually overlapping affective states such as anger, fear, sadness, fatigue, discomfort, or neutral facial tension. The absence of physiological, behavioral, contextual, and self-reported information further restricts the model's ability to determine whether a facial expression reflects stress or another emotional condition. Multimodal integration may reduce this ambiguity, though it would introduce additional sensor, synchronization, privacy, and computational requirements.

Fifth, while Grad-CAM improves visual transparency by highlighting salient facial regions, it remains a post-hoc, approximate attribution method that does not establish causal physiological mechanisms or clinically validated facial biomarkers of stress. Collectively, these limitations restrict the external validity and clinical applicability of the proposed framework. The system should therefore be regarded as a lightweight, interpretable, and reproducible research baseline for stress-related facial-expression proxy classification rather than as a clinical stress detector or diagnostic tool. Before practical or clinical deployment, the framework must be evaluated using independently collected real-world cohorts, clinically or physiologically grounded stress labels, spontaneous temporal data, demographic fairness analysis, external validation, model-calibration assessment, and prospective comparison with established stress-assessment procedures.

VI. Conclusion

This study presented a lightweight and interpretable convolutional neural network (CNN) framework with Grad-CAM for noninvasive binary classification of stress-related and non-stress facial expressions on the FER2013 and CK+ benchmark datasets. The proposed model (0.42 million parameters, 0.03 GFLOPs) achieved 89.6% accuracy (89.6% F1-score, 0.93 AUC) on FER2013 and 89.8 ± 2.7% accuracy (90.2 ± 2.6% F1-score) under subject-independent leave-one-subject-out validation on CK+, outperforming VGG16, ResNet50, and EfficientNet-B0 in predictive balance while reducing inference latency to 12.0 ± 0.8 ms on an NVIDIA RTX 4060 GPU. Grad-CAM visualizations and deletion-based faithfulness analysis confirmed that predictions were driven primarily by expression-relevant facial regions rather than arbitrary background features. The principal contribution of this work is a computationally efficient, transparent, and reproducible research baseline for stress-related facial-expression proxy classification. Because the binary labels were derived from emotion annotations rather than clinically validated stress measurements, the framework should not be deployed as a clinical diagnostic tool. Future research will focus on validating the model on real-world, clinically annotated datasets across diverse demographic and environmental conditions,

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

incorporating temporal modeling and multimodal physiological signals, and optimizing edge-device deployment.

Abbreviations

AI: Artificial Intelligence
AUC: Area Under the Curve
BN: Batch Normalization
CK+: Extended Cohn–Kanade Dataset
CNN: Convolutional Neural Network
DL: Deep Learning
EEG: Electroencephalography
FER2013: Facial Expression Recognition 2013 Dataset
FLOPs / GFLOPs: Floating Point Operations / Giga Floating Point Operations
FPR: False Positive Rate
FPS: Frames Per Second
FP32: 32-bit Floating Point Precision
GPU: Graphics Processing Unit
Grad-CAM: Gradient-weighted Class Activation Mapping
IRB: Institutional Review Board
LOSO: Leave-One-Subject-Out
ReLU: Rectified Linear Unit
ROC: Receiver Operating Characteristic
SHAP: SHapley Additive exPlanations
XAI: Explainable Artificial Intelligence
YOLO: You Only Look Once

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no competing interests.

Ethical Approval

This study does not involve human participants or animals directly. All datasets used are publicly available and anonymized, and no ethical approval was required.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable. This study utilized publicly available, anonymized benchmark datasets (FER2013 and CK+) and did not involve any new human participants or individual identifiable data.

Data Availability

The datasets used in this study are publicly available from open repositories:

- The FER2013 dataset is available via Kaggle: (https://www.kaggle.com/datasets/msambare/fer2013)
- The Extended Cohn–Kanade (CK+) dataset is available at: (https://www.jeffcohn.net/resources/)

No new datasets were created during this study. Processed data and experimental results generated during the study are available from the corresponding author upon reasonable request.

Author Contribution

Sathish G and Priya Matta conceptualized and designed the study, handled data curation, and contributed to pipeline design and methodology. Anusha N and P. Pattabhirama Mohan contributed to model implementation, experimental evaluation, and drafting the manuscript. Aanandha Saravanan K and Dilli Ganesh V assisted with baseline validation, Grad-CAM interpretability analysis, and critical manuscript revisions. All authors reviewed and approved the final manuscript and agreed to be accountable for all aspects of the work.

Declaration of AI-Assisted Technologies

During the preparation of this manuscript, generative AI tools were used solely for language editing, readability enhancement, and grammar refinement in accordance with the journal's editorial policies. The authors reviewed and edited all content produced by these tools and take full responsibility for the final contents of the publication.

Acknowledgments

Institutional, technical, and administrative assistance provided by the participating institutions during the preparation of this study is gratefully acknowledged.

References

[1]. S. S. Shinde and A. S. Ghotkar, "Mental stress detection with the multimodal data using ensemble optimization enabled explainable convolutional neural network," *Biomedical Materials & Devices*, vol. 4, no. 1, pp. 693–715, Jan. 2026, doi: 10.1007/s44174-025-00192-x.

[2]. I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2015, doi: 10.1016/j.neunet.2014.09.005

[3]. Y. Xing, Y. Yang, K. Yang, A. Lu, L. Xing, K. Mackie, and F. Guo, "Intelligent sensing devices and systems for personalized mental health," *Med-X*, vol. 3, no. 1, Art. no. 10, Feb. 2025, doi: 10.1007/s44258-025-00057-3.

[4]. G. Giannakakis, A. Roussos, C. Andreou, S. Borgwardt, and A. I. Korda, "Stress recognition identifying relevant facial action units through explainable artificial intelligence and machine learning," *Computer Methods and Programs in Biomedicine*, vol. 259, Art. no. 108507, Feb.

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

2025, doi: 10.1016/j.cmpb.2025.108507.

[5]. S. Bouazizi and H. Ltifi, "Emotion recognition with EEG signals using an insightful adapted echo state network framework," *International Journal of Information Technology*, vol. 17, no. 2, pp. 1–11, Mar. 2025, doi: 10.1007/s41870-025-02850-0.

[6]. M. Z. Naveed and D. S. Aparna, "Facedepai: A deep learning framework for depression detection from facial expressions using Facedepnet," *Journal of Theoretical and Applied Information Technology*, vol. 103, no. 23, pp. 8320–8332, Dec. 2025

[7]. T. R. Lanjewar, A. Dutonde, and A. Yadav, "Real-Time Emotion Detection and Mental Health Monitoring Using YOLOv11 with Explainable AI," in *Proc. 4th OPJU Int. Technol. Conf. (OTCON)*, Raigarh, India, 2025, pp. 1–7, doi: 10.1109/OTCON61756.2025.10543210.

[8]. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, San Francisco, CA, USA, 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.

[9]. V. Kukreja, "A novel multimodal framework for emotion recognition in comics: Integrating background color, facial expressions, and text sentiment using fuzzy logic," *Expert Systems with Applications*, vol. 289, Art. no. 128267, Mar. 2025, doi: 10.1016/j.eswa.2025.128267.

[10]. F. Yang and F. Gong, "A multimodal spatiotemporal convolutional network with attention mechanism for athlete anxiety behavior recognition," *Scientific Reports*, vol. 16, Art. no. 3142, Feb. 2026, doi: 10.1038/s41598-026-03142-8

[11]. L. Shi, "Exploring students' emotion recognition and teachers' teaching feedback in college foreign language classroom based on AFCNN model," *Scientific Reports*, vol. 16, Art. no. 1824, Jan. 2026, doi: 10.1038/s41598-026-01824-3.

[12]. W. Li, D. Huang, H. Li, and Y. Wang, "Automatic 4D facial expression recognition using dynamic geometrical image network," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Xi'an, China, 2018, pp. 24–30, doi: 10.1109/FG.2018.00013.

[13]. Gandhi, V., Chaudhari, Y., Kumar, A. et al. Benchmarking Machine Learning Models for Obesity Classification with SHAP-Based Interpretability. Int J Comput Intell Syst (2025). https://doi.org/10.1007/s44196-025-01078-x

[14]. S. M. Saleem, S. R. Zeebaree, and M. B. Abdulrazzaq, "Real-life dynamic facial expression recognition: A review," *Journal of Physics: Conference Series*, vol. 1963, no. 1, Art. no. 012010, Jul. 2021, doi: 10.1088/1742-6596/1963/1/012010.

[15]. M. G. Garcés, J. C. Montoya, M. I. P. Martínez, J. V. García, and E. H. H. Rincón, "Deep learning and noninvasive sensors for detecting physiological dysregulation: A scoping review," *Journal of Medical Systems*, vol. 50, no. 1, Art. no. 14, Jan. 2026, doi: 10.1007/s10916-025-02114-1.

[16]. N. C. Kansara, B. Vala, and M. S. Shaikh, "Leveraging AI for detection of mental health disorders: An investigation of multiple classifiers," *IET Conference Proceedings*, vol. CP920, no. 7, pp. 419–426, May 2025, doi: 10.1049/icp.2025.0419.

[17]. H. S. Khurana and P. D. Joshi, "Deep learning-based stress detection using facial expression recognition and the AffectNet dataset," in *Smart Trends for Information Technology and Computer Communications*, CCIS vol. 2145. Singapore: Springer, 2025, pp. 425–433, doi: 10.1007/978-981-97-8910-4_32.

[18]. F. B. Kunwar, M. Saraswat, J. Singh, R. Prameela Devi, S. J. Mohan, and P. S. Bhati, "Emotion recognition using deep learning on EEG data for stress and anxiety detection," in *Advanced Computing and Intelligent Technologies*, CCIS vol. 2110. Singapore: Springer, 2024, pp. 233–245, doi: 10.1007/978-981-97-6540-1_18.

[19]. B. V. Gokulnath, P. Charmitha, P. Chathurya, D. L. Satya Sri, B. Vennela, S. S. Ibrahim, and S. S. Kumar, "Empowering emotional intelligence through deep learning techniques," *Scientific Reports*, vol. 15, Art. no. 21850, Aug. 2025, doi: 10.1038/s41598-025-21850-7.

[20]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[21]. Y. Yaddaden, "Efficient dynamic emotion recognition from facial expressions using statistical spatio-temporal geometric features," *Big Data and Cognitive Computing*, vol. 9, no. 8,

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Art. no. 213, Aug. 2025, doi: 10.3390/bdcc9080213.

[22]. F. Gong, "Design and implementation of an intelligent educational interaction system with integrated multimodal emotion recognition and adaptive content delivery," *Discover Artificial Intelligence*, vol. 6, no. 1, Art. no. 48, Jan. 2026, doi: 10.1007/s44163-026-00048-w.

[23]. A. Chauhan and S. Vegad, "CVAE-ADS: A deep learning framework for traffic accident detection and video summarization," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 8, no. 1, pp. 185–205, Jan. 2026, doi: 10.35882/jeeemi.v8i1.1139.

[24]. X. Wu, T. Lee, U. K. Lilhore, S. Simaiya, R. Alroobaea, A. M. Baqasah, and L. G. Tekeste, "A deep learning approach to emotionally intelligent AI for improved learning outcomes," *Scientific Reports*, vol. 16, Art. no. 4521, Mar. 2026, doi: 10.1038/s41598-026-04521-1

[25]. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019, pp. 6105–6114.

[26]. H. Zhang, Z. Pang, P. Zhao, G. Tang, L. Shen, and G. Wang, "Pre-attentive speech signal processing with adaptive routing for emotion recognition," *Biomedical Signal Processing and Control*, vol. 112, Art. no. 108782, Feb. 2026, doi: 10.1016/j.bspc.2025.108782.

[27]. T. Gu, H. Li, X. Feng, and Y. Luo, "AMGSN: Adaptive mask-guide supervised network for debiased facial expression recognition," *Pattern Recognition*, vol. 170, Art. no. 112023, Feb. 2026, doi: 10.1016/j.patcog.2025.112023.

[28]. K. Wang, D. Wu, W. Liu, H. He, and H. Chen, "Graph filtering-based multi-granularity spatial-frequency convolutional attention fusion network for emotion recognition," *Biomedical Signal Processing and Control*, vol. 112, Art. no. 108556, Feb. 2026, doi: 10.1016/j.bspc.2025.108556..

[29]. W. Wang, D. Yang, Y. Yang, Y. Xie, X. Liu, Y. Yu, and K. Shi, "CV-EEGNet: A compact complex-valued convolutional network for end-to-end EEG-based emotion recognition," *Sensors*, vol. 26, no. 3, Art. no. 807, Jan. 2026, doi: 10.3390/s26030807.

[30]. Shehada, D., Turky, A., Khan, W., Khan, B., & Hussain, A. (2023). A lightweight facial emotion recognition system using partial transfer learning for visually impaired people. *IEEE Access, 11*, 36961–36969. https://doi.org/10.1109/ACCESS.2023.3264268

[31]. N. Zhou, L. Jika, W. Wang, Y. Xie, Y. Du, and Y. C. Soh, "BERN: A novel framework for enhanced emotion recognition through the integration of EEG and eye movement features," IEEE Transactions on Cognitive and Developmental Systems, vol. 17, no. 3, pp. 589–601, Sep. 2025, doi: 10.1109/TCDS.2025.3382109

[32]. H. Tan, X. Lu, C. Zhou, Y. Si, J. Shi, D. Sun, and H. Zhang, "Human emotion decoding for Mandarin with anti-background noise mask and deep learning," *ACS Applied Nano Materials*, vol. 8, no. 36, pp. 17610–17619, Sep. 2025, doi: 10.1021/acsanm.5c02418.

[33]. Y. Li and L. Peng, "Emotion-aware mental health intervention strategies on mobile computing platforms," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 17, pp. 112–126, Sep. 2025, doi: 10.3991/ijim.v19i17.41285.

[34]. V. C. Gandhi and P. P. Gandhi, "A Survey - Insights of ML and DL in Health Domain," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2022, pp. 239-246, doi: 10.1109/ICSCDS53736.2022.9760981.

[35]. L. Giles and J. E. Khoury, "Emotion regulation during pregnancy: A pathway from maternal childhood maltreatment to perinatal mental health," *Journal of Affective Disorders*, vol. 385, Art. no. 119417, May 2025, doi: 10.1016/j.jad.2025.119417.

[36]. E. Zangeneh and A. Moradi, "Facial expression recognition by using differential geometric features," *The Imaging Science Journal*, vol. 66, no. 8, pp. 463–470, Dec. 2018, doi: 10.1080/13682199.2018.1517036.

[37]. R. Yan, M. Yang, Q. Zheng, D. Wang, and C. Peng, "Facial expression recognition based on hybrid geometry-appearance and dynamic-still feature fusion," *Multimedia Tools and Applications*, vol. 82, no. 2, pp. 2663–2688, Jan. 2023, doi: 10.1007/s11042-022-13248-0.

[38]. Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. Sensors, 21(9), 3046. https://doi.org/10.3390/s21093046C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, article 60, 2019, doi:

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

10.1186/s40537-019-0197-0.
[39]. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[40]. V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, 2010, pp. 807–814.
[41]. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, 2015, pp. 448–456.
[42]. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.
[43]. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, 2015.
[44]. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.
[45]. T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010
[46]. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

Author Biography



Dr. Sathish Gajendran, who holds a Ph.D. in Electronics and Communication Engineering, is currently serving as an Assistant Professor in the School of Computing at the Bharath Institute of Science and Technology. He brings over 20 years of combined experience spanning academic teaching, scientific research, and industry practice. An active member of the International Association of Engineers (IAENG) Society of Electrical Engineering and Wireless Networks, his primary research focuses on carbon nanotube field-effect transistors (CNTFETs) with high-K dielectrics, low-power electronics, quantum capacitance modeling, and drain current optimization for next-generation energy-efficient semiconductor devices.



Dr. Priya Matta is currently serving as Associate Professor and Head of the Department of Computer Applications at Tula's Institute, Dehradun, India. She possesses over 21 years of experience across teaching, research, and academic leadership. She earned her Ph.D. in Computer Science and Engineering from Graphic Era University, alongside an M.Tech. as a Gold Medalist and an MCA. Her research focuses on Cyber Security, Data Science, AI/ML, and IoT. Dr. Matta has authored more than 80 papers in reputed international journals and conferences. A Senior Member of IEEE, she actively contributes to ACM, ISTE, and CSI.



Mrs. Anusha N. is currently working as an Assistant Professor in the Department of Information Science and Engineering at NMAM Institute of Technology, Karnataka, India, with over nine years of dedicated teaching experience. She earned her B.E. and M.Tech. in Computer Science and Engineering from NMAMIT, affiliated with Visvesvaraya Technological University, Belagavi. Currently, she is pursuing her Ph.D. in Computer Science and Engineering at Nitte Deemed to be University. Her research interests include deepfake detection, machine learning, and artificial intelligence in healthcare and forensics. She has authored ten research papers and holds active memberships in IEEE and ISTE.



Pattabhirama Mohan Patnala is currently serving as an Assistant Professor in the Department of Information Technology and Data Science at Aditya University, while also pursuing his Ph.D. His primary research focuses on automated brain tumor detection and classification using advanced deep learning techniques applied to MRI scans. Beyond neuro-imaging, his broader academic interests encompass medical image processing, convolutional neural network (CNN) architectures such as Inception V3, and the development of intelligent healthcare systems. Additionally, he has made notable research contributions to the Internet of Things (IoT) and modern communication technologies, aiming to bridge artificial intelligence with practical clinical and smart-system applications.



Dr. K. Aanandha Saravanan is currently working as Professor in Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai. He received M.E (Communication Systems) in 2006 from

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Noorul Islam College of Engineering, Nagercoil Tamil Nadu and PhD in 2022 from VelTech, India. He has a teaching experience of about 26 years in various engineering colleges. His research interest includes Artificial Intelligence, Machine Learning, Data Science, Internet of Things, Health Care, Network Security and Wireless Communication Networks. He has published more than 80 papers in reputed Journals and Conferences.



Dr. Dilli Ganesh V is an Assistant Professor at Saveetha Institute of Medical and Technical Sciences. He is recognized among the World's Top 2% Scientists in 2025 for his significant contributions to research and innovation. His research interests include Artificial Intelligence, Internet of Things, Wireless Sensor Networks, Blockchain, and emerging smart technologies. Dr. Dilli Ganesh V has published numerous high-impact research articles in reputed international journals and conferences. He has also authored books, filed patents, and contributed extensively to academic and industrial research activities. His work focuses on developing innovative and sustainable technological solutions for real-world applications.

Manuscript received 10 July 2026; Accepted 25 September 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.1918
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).