Computing and Artificial Intelligence 2025, 3(4), 4397.
doi: https://doi.org/10.59400/cai4397

Article

Leveraging extensive feature modeling for facial emotion recognition

Milica Tufegdzic[1,*], Nevena Tufegdzic[2], Marija Mojsilovic[1]

[1] Academy of Professional Studies Sumadija, Kragujevac 34000, Serbia
[2] SaTCIP Publisher Ltd, Vrnjacka Banja 36210, Serbia
*** Corresponding author:** Milica Tufegdzic, mtufegdzic@asss.edu.rs

CITATION

Tufegdzic M, Tufegdzic N, Mojsilovic M. Leveraging extensive feature modeling for facial emotion recognition. Computing and Artificial Intelligence. 2025; 3(4): 4397. https://doi.org/10.59400/cai4397

ARTICLE INFO

Received: 13 October 2025
Revised: 5 December 2025
Accepted: 8 December 2025
Available online: 19 December 2025

Abstract: Facial emotion recognition (FER) is an important area of affective computing with applications in human–computer interaction, healthcare, education, and intelligent systems. Although recent FER research is largely dominated by deep learning and transformer-based approaches, handcrafted feature modeling remains attractive due to its interpretability and lower computational requirements. This study proposes an Action Unit–based machine learning (AU-ML) framework for recognizing basic emotions from lateral facial expressions using the Karolinska Directed Emotional Faces (KDEF) dataset. Facial Action Units (AUs) were extracted, and manual feature selection was performed to retain only AU intensity and presence information relevant to emotion recognition. This process significantly reduced the original feature vector, improving computational efficiency while preserving classification performance. To compensate for the reduced dataset size after extracting lateral images, data augmentation techniques, including horizontal flipping, shifting, scaling, rotation, and brightness and contrast adjustments, were applied prior to AU extraction. Several machine learning algorithms were evaluated, including K-Nearest Neighbors, Support Vector Classifier, Decision Tree, Naïve Bayes, Random Forest, AdaBoost, Bagging, Voting, and Stacking Classifiers, CatBoost, and Extreme Gradient Boosting. The results demonstrated that ensemble methods generally outperformed simpler classifiers, with CatBoost achieving the best classification performance. The findings indicate that extensive feature modeling remains a reliable approach to emotion recognition, and that AU-based representations provide an interpretable and computationally efficient alternative to deep learning approaches.

Keywords: machine learning; action units; emotion classification; image processing; feature extraction; FACS (facial action coding system); explainable AI

1. Introduction

With the recent expansion of artificial intelligence (AI) systems, the general public is progressively becoming more reliant on AI agents for both everyday and professional tasks. These complex agents interpret the input, be it text, image, or voice, and use generative AI to produce an output in response to the initial prompt. Modern research is heavily focused on the various possibilities of these agents, utilizing transformer networks to produce high-quality results in real-time. The interpretation stage is now heavily intertwined with the generative stage, as modern transformer systems do both of these tasks efficiently [1]. Traditional and deep learning-based approaches mainly differ in how their performance scales with the amount of data [2], but traditional tasks like regression and classification are still as important as ever. Their usage in

complex systems is undeniably present, from the application of computer vision (CV) in self-driving cars to analyzing big data for business intelligence and decision-making systems, where machine learning (ML) and deep learning (DL) models do best.

One such instance where traditional AI models still find a purpose is emotion recognition (ER). While these models are not to be used individually, integrating them into existing systems can be very beneficial for the end user. For example, a study by Xu et al. [3] explores the use of real-time emotion recognition systems based on convolutional neural networks (CNN) and long short-term memory (LSTM) models to monitor users' emotional states and dynamically adapt system behavior on an online learning platform. The results show that such adaptive systems significantly improve user engagement, reduce task completion time, and enhance overall satisfaction in human–computer interaction environments. This is but one application of emotion recognition, as present-day research aims to integrate such systems into medical applications and patient care [4, 5].

ER is a core task in affective computing, aiming to infer human emotional states from observable signals. Traditionally, ER has been approached through single modalities, such as text, speech, or visual cues, before evolving toward multimodal emotion recognition (MER), which integrates multiple sources of information for improved performance [6–8].

Visual emotion recognition relies on facial expressions, which are widely regarded as universal indicators of human emotional states. As such, the domain of this problem lies in CV and image processing. Early approaches utilized handcrafted features, such as Local Binary Patterns (LBP) and Histograms of Oriented Gradients (HOG), to capture facial characteristics [9]. More recently, deep learning methods such as CNNs have become dominant, enabling the automatic extraction of complex features from facial data. This modality offers the advantage of directly observing emotional expressions and provides strong cues for clearly distinguishable emotions such as happiness, anger, and surprise. However, its effectiveness is influenced by several factors, including occlusion, lighting conditions, and variations in head pose, which can degrade performance. Additionally, cultural differences may affect how facial expressions are interpreted, and certain emotions, such as sadness, can be more subtle and less visually distinct, which presents a challenge for accurate recognition [10].

There are two main approaches in FER:

1. Utilizing hand-crafted feature extraction in combination with ML or DL models;
2. Utilizing pure DL and transfer learning.

These paradigms differ primarily in how features are obtained and represented, which directly impacts model performance, interpretability, and robustness.

Early FER systems rely on explicitly designed feature descriptors that aim to capture various facial characteristics, such as texture, shape, and local intensity variations. These features are then fed into classifiers such as Support Vector Machines (SVMs) or shallow neural networks. A key prerequisite for these methods is accurate face detection and preprocessing, as many algorithms require a normalized facial region extracted from the original image. One of the most widely used algorithms are texture-based descriptors such as:

1. Haar-like features combined with cascade classifiers (e.g., Viola–Jones) [11,12];
2. HOG [9,13];
3. LBP [11,14,15].

Haar-based methods detect contrast patterns (edges, lines, and regions) and use boosting techniques such as AdaBoost to construct strong classifiers from weak learners [16]. While computationally efficient, these methods are sensitive to variations in pose, illumination, and occlusion [11].

HOG descriptors represent local object structure by analyzing gradient orientations, making them more robust to illumination changes. However, they often fail to capture fine-grained facial details essential for distinguishing subtle emotions [13]. This limitation could be addressed by selecting a proper set of parameters, which results in generating more discriminative features [17]. **Figure 1** represents an example of feature extraction using HOG descriptors. The example images strongly suggest that while HOG captures coarse information with ease, such as eye and mouth position, fine details such as furrowed brows are easily missed [11].

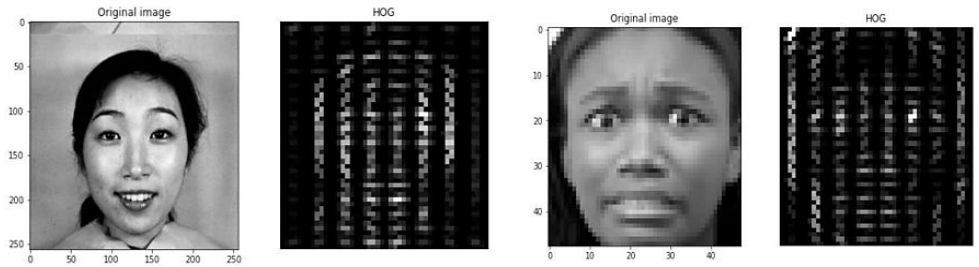

Figure 1. HOG descriptor on human faces [9].

LBP encodes local texture by comparing pixel intensities within a neighborhood, producing compact and efficient representations. Despite their robustness to illumination changes and suitability for low-resolution images, LBP-based methods are highly sensitive to noise [18].

Additional feature extraction techniques include Gabor filters, which analyze spatial-frequency information and are particularly effective for detecting edges and fine facial structures [19], and color histograms, which capture chromatic variations. Gabor filters operate by convolving a complex exponential function with a 2D Gaussian kernel, producing responses at different scales and orientations. These responses generate Gabor images that encode local spatial-frequency and orientation information [20]. Gabor filters have limited applicability in FER due to minimal color changes across emotional states [19]; however, they are especially valuable due to their alignment with human visual perception, enabling effective representation of facial textures and micro-patterns. Finally, they are computationally expensive and sensitive to noise.

Another important class of handcrafted methods focuses on facial geometry, modeling the spatial configuration of key facial landmarks. These approaches typically involve the detection of facial landmarks (e.g., eyes, nose, mouth, jawline), construction of feature vectors based on distances or angles between landmarks and analysis of facial action units (AUs) corresponding to muscle movements. Datasets such as Multi-PIE, AFLW, and HELEN provide annotated landmarks used to train such

models [21–23]. For example, a common representation involves computing Euclidean distances between selected landmark pairs to capture emotion-induced deformations.

Geometric methods are interpretable and computationally efficient, making them suitable for real-time applications. However, their performance heavily depends on accurate landmark detection and may degrade under pose variations or occlusions.

Hand-crafted features are typically coupled with traditional classifiers such as:

1. SVMs [24];
2. Linear Discriminant Analysis (LDA) (Fisherfaces) [25];
3. Principal Component Analysis (PCA) (Eigenfaces) [26].

SVMs provide strong generalization capabilities, especially when combined with feature descriptors such as HOG [24]. However, sliding-window approaches used in detection can be computationally inefficient and struggle with occlusions. LDA maximizes class separability, making it more suitable for classification tasks [27]. Nevertheless, it assumes linear separability of classes, which is often not applicable in real-world FER scenarios. In contrast, PCA-based approaches reduce dimensionality by projecting facial data onto directions of maximum variance, enabling compact representations, but they are sensitive to illumination and alignment issues [26].

With the advancement of computational resources and large-scale datasets, deep learning has become the dominant paradigm in FER. Unlike traditional methods, convolutional deep neural networks automatically learn hierarchical feature representations directly from raw pixel data, eliminating the need for manual feature engineering. In other words, CNN-based models learn progressively abstract features in order [28]:

1. Low-level features (edges, textures);
2. Mid-level features (facial components);
3. High-level features (semantic representations of expressions).

This hierarchical learning enables improved robustness to variations in pose, illumination, and occlusion. For those reasons, CNN architectures such as ResNet, EfficientNet, and DenseNet are commonly used for FER tasks [29]. Transfer learning plays a crucial role, where models pretrained on large datasets (e.g., ImageNet) are fine-tuned for emotion recognition. This approach offers several advantages, such as reduced training time, improved generalization on limited datasets, and better feature representations. As a result, modern FER pipelines often rely on advanced face detection models such as Multi-task Cascaded CNN (MTCNN) [30], You Only Look Once (YOLO) [31], and RetinaFace [32]. MTCNN employs a cascade of networks (P-Net, R-Net, O-Net) to progressively refine face detection and landmark localization. YOLO enables real-time detection by processing the entire image in a single forward pass. RetinaFace leverages feature pyramid networks and backbone architectures such as ResNet50 for highly accurate detection, even for small faces.

While hand-crafted methods offer interpretability and lower computational cost, they are limited in their ability to generalize across real-world conditions. Deep learning approaches, although more accurate, require large datasets and significant computational resources. Key challenges in FER remain: sensitivity to occlusion and

head pose, variability in lighting and image quality, cultural and individual differences in emotional expression, and difficulty in recognizing subtle or compound emotions. DL-based models are able to differentiate emotions with better accuracy, but offer significantly less in terms of their interpretability.

This study presents a systematic comparison of multiple machine learning algorithms applied to a compact AU-based embedding and evaluates their performance. Additionally, the proposed approach emphasizes interpretability by explicitly linking classification outcomes to the most relevant AUs.

Although AU-based FER has been previously investigated, this study differs from prior work in several important aspects. The study introduces a compact and interpretable AU embedding obtained through domain-driven feature selection from high-dimensional OpenFace outputs. An extensive benchmark of conventional and ensemble machine learning methods is performed to evaluate the effectiveness of interpretable AU-based representations. Finally, the study presents a prototype explainable FER application suitable for computationally constrained environments.

2. Materials and methods

Prior to presenting the feature extraction and classification methods, the problem domain must be established. Humans have a wide array of emotions that can be stronger or weaker. Another factor is cultural and age differences, which can also influence how humans express emotion. No two individuals experience emotions identically, highlighting the role of intrapersonal differences. For these reasons, studies on facial emotion recognition usually reference a subset of basic emotions [33], which consists of the following: joy, surprise, fear, disgust, sadness, and anger. Since these emotions are universal and easy to recognize, they will be used as a benchmark for the proposed approach.

The purpose of this study is to demonstrate how utilizing domain knowledge yields better results compared to pure DL or transformer models. Handcrafted features are easier to interpret and contain valuable information that can later be used for various purposes and not just for classifying the emotions on the face. AUs are a prime example of this approach since they directly translate domain knowledge, in this case psychology and human anatomy, into quantifiable data. By extracting the key AUs on the face, it is possible to derive a handcrafted embedding vector that can later be fed into an ML model. The ML model then performs a classification, and the results can be visualized and explained directly by analyzing the key AUs that led to such classification. AUs are enumerated and divided into groups, totaling 46 AUs on the human face [34]; however, for this study, only the AUs that form expressions for basic emotions are of importance. **Table 1** shows how different AUs are activated for different basic emotions.

Table 1. AUs per emotion.

Emotion	Action units
Joy	6 + 12 + 25
Surprise	1 + 2 + 5 + 25 + 26
Fear	1 + 5 + 11 + 20 + 25 + 26

Table 1. *Cont.*

Emotion	Action units
Disgust	4 + 6 + 9 + 11 + 15 + 17
Sadness	1 + 4 + 15
Anger	4 + 5 + 23 + 38

As shown in **Table 1**, only a subset of 15 AUs is used to form a basic emotion expression. The remaining AUs are not relevant to this study and will therefore be excluded from the construction of the facial embedding vector.

In order to apply the AU system, a dataset containing images of front-facing humans is necessary. Ideally, the images should be high resolution and grayscale, with sufficient diversity across age and sex. One such dataset is the Karolinska Directed Emotional Faces (KDEF), which contains 7 emotions (6 basic emotions along with a neutral class) across 4,900 samples. This dataset was derived under laboratory conditions by recording 272 participants in front of a camera while capturing both frontal and lateral facial expressions [35]. While the frontal images are perfect for AU detection, the lateral ones present a challenge that many FER models face, which is the inability to correctly determine the features of the face. As a consequence, lateral FER is considered a problem by itself and presents a possibility for future research. In order to use KDEF, it is necessary to extract all lateral images, which was performed manually in his study.

After extracting the lateral images, the dataset was reduced to approximately one third of its original size. To increase the number of samples, several data augmentation techniques were applied, including horizontal flipping, shifting, scaling, rotation, and random adjustments in brightness and contrast. During the augmentation process, no alterations were introduced to facial proportions, as the aspect ratio of the images was preserved. As a result, the dataset was expanded to 7,800 samples, providing a suitable basis for benchmarking the AU-ML approach proposed in this study. The distribution of samples per emotion is depicted in **Figure 2**.

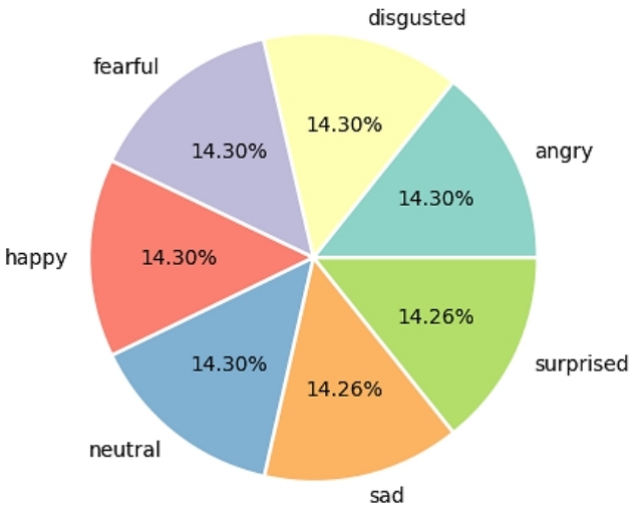

Figure 2. KDEF distribution of samples per emotion.

A library called OpenFace is used in order to capture AUs on faces. OpenFace is

a system designed to perform various tasks tied to FER, such as detecting faces in an image, detecting the pose of each face, key facial landmarks, the eye and gaze position, as well as the detection of AUs themselves [36]. The AUs that OpenFace detects are presented in **Table 2**.

Table 2. OpenFace AUs.

Au	Full name	Prediction
AU1	Inner Brow Raiser	I
AU2	Outer Brow Raiser	I
AU4	Brow Lowerer	I
AU5	Upper Lid Raiser	I
AU6	Cheek Raiser	I
AU7	Lid Tightener	P
AU9	Nose Wrinkler	I
AU10	Upper Lip Raiser	I
AU12	Lip Corner Puller	I
AU14	Dimpler	I
AU15	Lip Corner Depressor	I
AU17	Chin Raiser	I
AU20	Lip Stretcher	I
AU23	Lip Tightener	P
AU25	Lips Part	I
AU26	Jaw Drop	I
AU28	Lip Suck	P
AU45	Blink	P

Source: Adapted from Barrett et al. [33].

According to **Table 2**, OpenFace detects more than the 15 AUs listed above, with the additional AUs being AU7, AU10, AU14, AU28, and AU45. On the other hand, the detection of AU11 (Nasolabial Deepener) and AU38 (Nostril Dilator) is missing; however, these are not considered crucial AUs for emotion recognition. According to the Facial Action Coding System (FACS) specification, AU6 is characterized by cheek raising and eyelid compression. In the OpenFace implementation, this action is separated into AU6 and AU7, and therefore AU7 will also be taken into consideration during the implementation of the proposed solution.

OpenFace first detects facial landmarks by using the Constrained Local Neural Fields algorithm, which is a patch-based algorithm based on computing local neural fields [37]. Each landmark is first detected by a local detector, which is followed by a global facial landmark alignment function based on a PCA model trained on pre-annotated datasets for facial landmark detection. For local neural fields, a HOG descriptor is used to determine the exact landmark position inside a small region of pixels, 15×15 in size. The result of this detection is 68 facial landmarks, which are further used to determine other OpenFace outputs [37].

The facial landmark output is then compared to a neutral face representation, and differences are then coupled with raw CLNF results to form the input for AU detection. This allows for accurate and explainable AU detection tested on multiple datasets such as FER-2013, CK+, KDEF, and RAF-DB [37]. **Figure 3** shows a visualized OpenFace output on a sample image taken from the KDEF dataset. The figure shows facial

landmarks (red), gaze direction (green), and the facial bounding box, which additionally represents facial pose (blue).

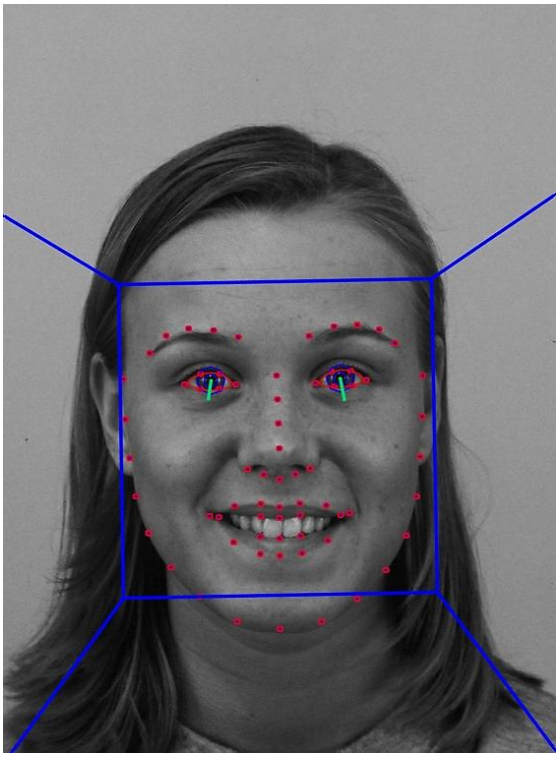

Figure 3. OpenFace results on a KDEF sample.

3. Results

The output of AU detection is a feature vector containing more than 700 dimensions. In addition to AU-related information, the vector includes numerous auxiliary byproducts of the detection process, such as facial landmark coordinates, facial bounding box coordinates, head pose estimation parameters, gaze direction, and the position of the eyeball sphere in the estimated 3D image space. Empirical evaluation showed that these features did not improve classification performance, but instead increased the computational complexity and response time of the system. Therefore, feature selection was performed to retain only AU intensity and presence values relevant to emotion recognition. The selected features included AUs associated with the formation of basic emotional expressions, as well as AUs for which only presence information could be detected. Following this manual dimensionality reduction process, the final feature vector was reduced to 34 dimensions.

After processing the KDEF dataset and removing failed detections, a total of 7,787 samples remained. Considering that the original, unprocessed dataset contained 7,900 samples, this represents an effective and reliable feature extraction. Failed detections refer to images in which facial detection and subsequent processing could not be successfully performed, mostly due to head pose or occlusions. However, because the failed samples were not distributed uniformly across emotion categories, the exclusion process resulted in a slight class imbalance in the final dataset. This imbalance was taken into consideration during model selection and evaluation, particularly for classifiers sensitive to uneven class distributions. Before evaluating the performance of

different ML models, the dataset was split into training and testing sets using a 4:1 ratio with a fixed random seed for selection. Consequently, 6,229 randomly selected samples were assigned to the training set, whereas the remaining 1,558 samples comprised the testing set.

All the models were tested using the Python programming language, since it provides a vast ecosystem of modules suited for ML. The training and inference were conducted on an Intel i5 12600K machine with an NVIDIA GeForce RTX 4060, with 16GB of available DDR5 RAM.

The following models were chosen for experimental evaluation: AdaBoost, Bagging Classifier, CatBoost, Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes, Random Forest, Stacking Classifier, Support Vector Classifier (SVC), Voting Classifier, and Extreme Gradient Boosting (XGBoost). The selected models comprise a combination of conventional classification algorithms and ensemble learning methods. In the case of the Stacking and Voting Classifiers, Decision Tree, KNN, Naïve Bayes, and SVC were employed as base estimators. Model hyperparameters were selected empirically to balance classification performance and computational efficiency.

The K-Nearest Neighbors (KNN) classifier was configured with k = 7 neighbors and the kd-tree search algorithm to improve computational efficiency during nearest-neighbor retrieval. The Support Vector Classifier (SVC) employed a radial basis function (RBF) kernel to model non-linear decision boundaries. Given the slight class imbalance introduced during preprocessing, balanced class weighting was enabled to reduce potential bias toward majority classes, while tie-breaking between competing class predictions was explicitly permitted. The maximum number of iterations was limited to 1,000 to constrain computational cost. The Naïve Bayes model was implemented using the Complement Naïve Bayes variant, which has been shown to exhibit greater robustness to imbalanced class distributions compared to standard Naïve Bayes. The Decision Tree classifier served as a simple, interpretable baseline model capable of learning hierarchical decision rules directly from the AU feature space. Decision trees recursively partition the feature space according to criteria that maximize class separability, making their predictions relatively easy to interpret.

The Random Forest classifier was implemented as an ensemble of decision trees using parallel computation to improve training efficiency. The AdaBoost classifier was included as a boosting-based ensemble approach that iteratively emphasizes incorrectly classified samples during training. By combining multiple weak learners into a weighted ensemble, AdaBoost aims to improve predictive accuracy.

The Bagging Classifier was implemented as a bootstrap aggregation ensemble using parallel processing and a fixed random seed to ensure reproducibility. Bagging improves robustness by training multiple models on randomly sampled subsets of the training data and aggregating their predictions. The Voting Classifier combined the predictions of four base classifiers (Decision Tree, KNN, Naïve Bayes, and SVC) through an ensemble voting mechanism. This approach leverages the complementary strengths of multiple classifiers to improve prediction stability and reduce model-specific biases. The Stacking Classifier was also constructed using

Decision Tree, KNN, Naïve Bayes, and SVC as base classifiers. Unlike voting, stacking employs a meta-learning strategy in which predictions generated by base learners are used as input for a higher-level classifier. A five-fold cross-validation procedure was applied during training to reduce overfitting and improve generalization. The use of heterogeneous base classifiers in the Voting and Stacking ensembles was intended to improve robustness and reduce the influence of biases associated with any single learning paradigm, including sensitivity to minor class imbalance.

Gradient boosting methods such as CatBoost and XGBoost were included due to their strong performance on structured tabular data and their ability to model complex decision boundaries while remaining robust to some imbalance in class distributions. The CatBoost classifier was configured to perform 5,000 training iterations with GPU acceleration enabled. The XGBoost classifier was implemented with 20 estimators, a maximum tree depth of 50, and a learning rate of 0.1. Multi-class classification was enabled through the multi: softmax objective function, while GPU acceleration was utilized to improve training performance.

Training time varied depending on the algorithm; however, in most cases, it did not exceed 2 s. The only exception was the CatBoost algorithm, for which the training process required 23.7 s.

Following training, the resulting models were evaluated on the testing dataset, and the corresponding classification report is provided in **Table 3**.

Table 3. Testing results.

Model	Accuracy	Precision	Recall	F1 score
AdaBoost	60.21%	61.02%	60.21%	55.04%
Bagging Classifier	93.33%	93.36%	93.33%	93.33%
CatBoost	96.98%	96.99%	96.98%	96.98%
Decision Tree	88.83%	88.98%	88.83%	88.87%
KNN	95.76%	95.79%	95.76%	95.74%
Naive Bayes	54.69%	68.09%	54.69%	46.43%
Random Forest	95.25%	95.27%	95.25%	95.25%
Stacking Classifier	95.76%	95.79%	95.76%	95.77%
SVC	91.91%	91.90%	91.91%	91.86%
Voting Classifier	94.35%	94.40%	94.35%	94.34%
XGBoost	94.42%	94.41%	94.42%	94.40%

The evaluation results are presented in the form of confusion matrices in **Figures 4–6**. The confusion matrix results presented in **Figure 4** show that the best-performing classifiers vary across emotions. For the Angry and Happy classes, both KNN and Random Forest achieved the highest accuracy (0.96 for Angry and 1.00 for Happy), while Random Forest performed best for Disgusted (0.97). KNN also showed strong performance for Fearful (0.95) and Sad (0.97), while Surprised was well recognized by KNN, Random Forest, and BaggingClassifier (0.93). In contrast, AdaBoost showed the weakest performance, with notable confusion between Angry and Sad (34%) and between Disgusted and Sad (26%), indicating misclassification in emotionally similar classes.

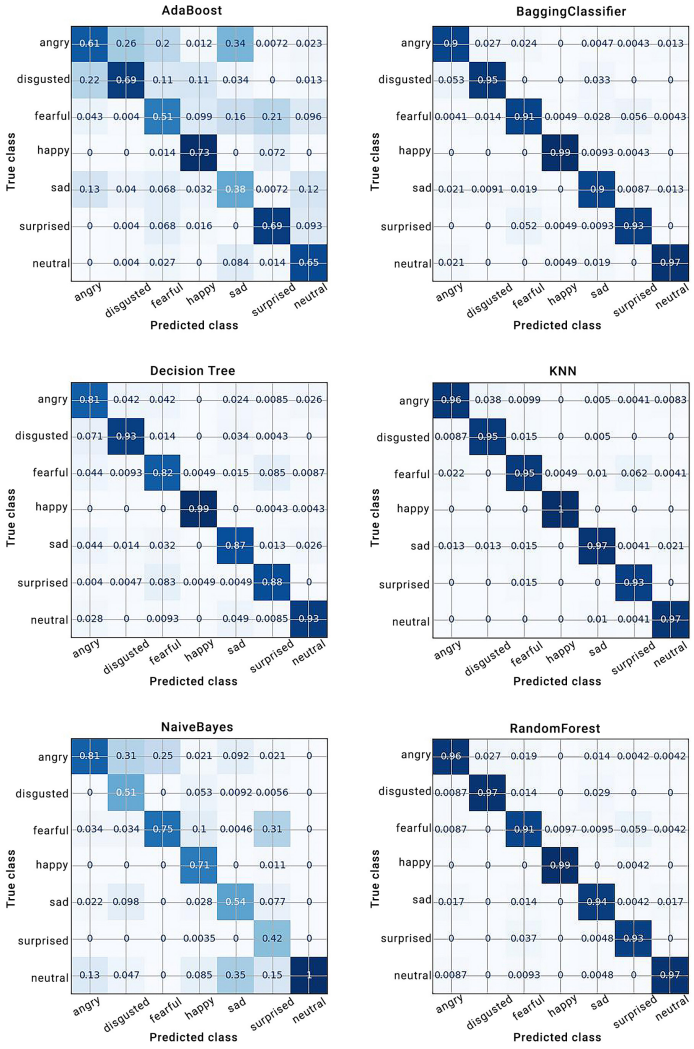

Figure 4. Confusion matrices for AdaBoost, BaggingClassifier, DecisionTree, KNN, Naive Bayes, and Random Forest.

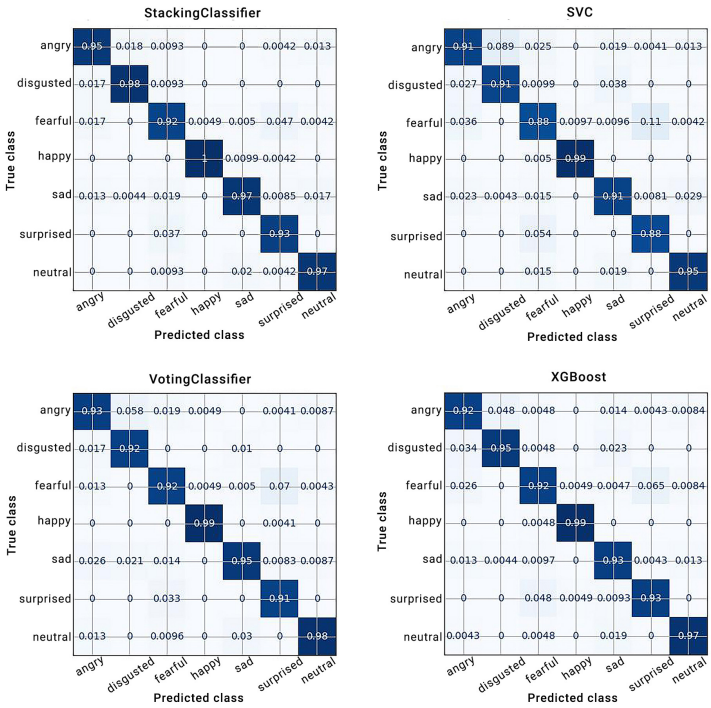

Figure 5. Confusion matrices for StackingClassifier, SVC, VotingClassifier, and XGBoost.

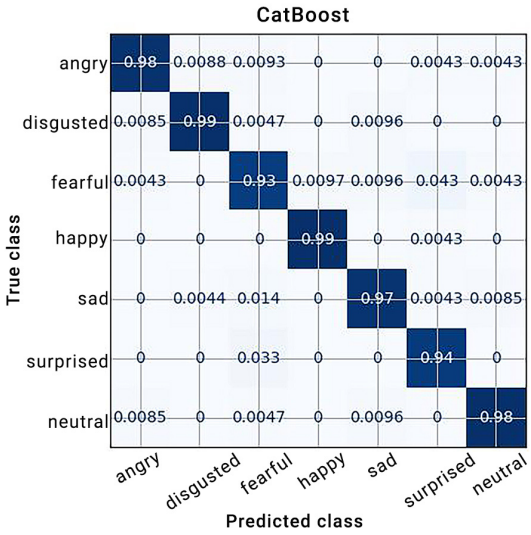

Figure 6. CatBoost confusion matrix.

The confusion matrix results presented in **Figure 5** indicate that the StackingClassifier achieved the best performance across most emotion classes, including Angry (0.95), Disgusted (0.98), Happy (1.00), and Sad (0.97), while also showing strong performance for Fear (0.92) and Surprised (0.93). For Fear, the VotingClassifier and XGBoost also achieved comparable performance (0.92), and for Surprised, XGBoost matched the StackingClassifier performance (0.93). Overall, the Happy class achieved perfect classification accuracy (1.00), indicating excellent model performance for this emotion category.

The results for CatBoost (**Figure 6**) show consistently strong performance across all emotion classes. The highest accuracy was achieved for Disgusted (0.99) and Happy (0.99), followed closely by Angry (0.98) and Sad (0.97), while Surprised (0.94) and Fear (0.93) showed slightly lower but still strong performance. Compared to previous results, the model demonstrates more balanced classification across all emotions, with improved consistency and reduced variation between classes.

In addition to model evaluation, a prototype application was developed to demonstrate the practical aspects of the proposed AU-ML framework in real-world scenarios. The application enables users to upload a facial image, after which the OpenFace toolkit is used to extract AUs and auxiliary facial analysis information. To improve interpretability, the interface visualizes OpenFace outputs, including facial landmarks, gaze direction, and estimated facial pose. Following feature extraction, the manually selected subset of AU intensity and presence values is processed according to the proposed dimensionality reduction strategy and subsequently passed to the trained CatBoost classifier for emotion recognition. The application outputs the predicted emotional category together with class probabilities, thereby providing insight into the confidence of the classification process. Furthermore, the five most prominent detected AUs are displayed, enabling interpretation of which facial muscle activations contributed most strongly to the predicted emotional state. **Figure 7** represents the interface of the described analysis inside the application.

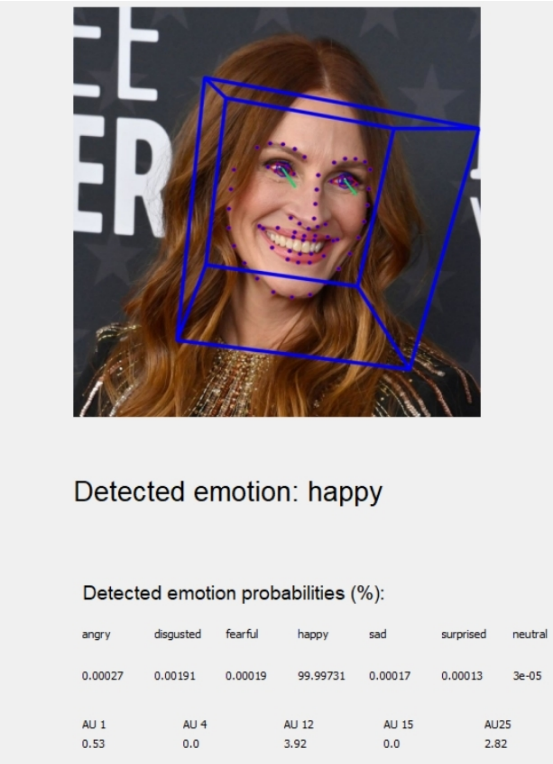

Figure 7. Visualized results of the proposed AU-ML approach.

This visualization demonstrates the explainability of the proposed AU-ML approach, representing an advantage over DL models commonly employed in FER. As model complexity increases, the interpretability of predictions generally decreases, making it more difficult to understand the decision-making process of deeper architectures. In contrast, the proposed framework provides transparency by exposing AUs that contribute to the final classification outcome. Furthermore, the use of conventional machine learning techniques reduces the computational requirements associated with both training and inference, making the approach suitable for deployment on low-end CPUs and potentially applicable in real-time environments.

4. Discussion

The results of this study demonstrate that facial AU-based feature representations can provide a robust basis for emotion recognition. The evaluated models achieved satisfactory classification performance, suggesting that a feature vector of AUs and their intensities represents a suitable approach for recognizing basic emotional states. This finding supports the underlying hypothesis that emotion recognition can be effectively performed using a reduced set of interpretable facial muscle activation patterns rather than relying solely on high-dimensional image representations.

The applied feature extraction strategy proved effective, as evidenced by the successful processing of 7,787 out of 7,900 images in the KDEF dataset. The relatively small number of failed detections indicates that the employed OpenFace-based pipeline demonstrated strong robustness in detecting facial landmarks and AUs under the given experimental conditions. Although the exclusion of unsuccessfully processed

samples introduced a slight class imbalance into the final dataset, this did not substantially impair classification performance. The use of classifiers capable of handling moderate imbalance, including the Support Vector Classifier (SVC) with balanced class weighting and ensemble-based methods such as Random Forest, XGBoost, and CatBoost, likely contributed to eliminating potential bias toward majority classes.

An important finding of this study concerns the effectiveness of manual dimensionality reduction. While OpenFace generates a feature vector exceeding 700 dimensions, empirical evaluation indicated that many auxiliary outputs, such as facial landmark coordinates, gaze direction, facial pose parameters, and bounding box information, did not significantly contribute to classification accuracy. Retaining only AU intensity and presence values relevant to emotional expression reduced the feature vector to 34 dimensions without degrading predictive performance. This finding aligns with previous studies emphasizing the importance of selecting physiologically meaningful features for facial expression recognition, rather than relying on large numbers of potentially redundant descriptors. By reducing feature dimensionality, the computational efficiency of the system was improved while preserving interpretability.

Based on the presented evaluation results, it is evident that the CatBoost model achieved the best classification performance. CatBoost is a boosting-based algorithm designed to improve predictive accuracy while reducing the risk of overfitting to the training dataset. It employs a gradient boosting approach in which an ensemble of weak learners, in this case decision trees, is trained sequentially to improve classification performance. Trees are constructed symmetrically, while the contribution of each tree to the final prediction is adaptively determined during training. This process is facilitated through adaptive learning rate optimization and the application of L2 regularization, which improves generalization and reduces overfitting [38]. The resulting ensemble model demonstrates strong classification performance while maintaining robustness against overfitting, making the obtained results consistent with expectations.

The comparative evaluation of machine learning models (**Table 3**) further demonstrated that ensemble approaches consistently outperformed simpler classification algorithms. While conventional classifiers such as Decision Tree, KNN, Naïve Bayes, and SVC provided acceptable results, the highest classification performance was achieved by CatBoost. This outcome is consistent with prior findings in structured-data classification, where gradient boosting methods frequently outperform individual learners due to their ability to model complex, non-linear interactions between features. The superior performance of CatBoost may be attributed to its gradient boosting framework, symmetric tree construction, adaptive learning mechanisms, and regularization strategies that improve generalization while reducing the risk of overfitting. The observed performance advantage suggests that AU-based emotion recognition benefits from ensemble methods capable of capturing subtle dependencies among facial muscle activations.

The effectiveness of ensemble learning was additionally reflected in the strong performance of Voting and Stacking Classifiers, which combined multiple base

estimators with complementary learning strategies. These findings indicate that no single classification paradigm fully captures the complexity of facial emotional expressions, and that combining multiple decision mechanisms can improve robustness. However, the consistently superior performance of CatBoost suggests that advanced boosting techniques may offer a more computationally efficient and accurate alternative to manually designed ensemble combinations.

The findings of this study should also be interpreted in the broader context of facial emotion recognition research. Many modern approaches rely heavily on deep convolutional neural networks trained directly on raw image data, often requiring vast computational resources and large-scale datasets while achieving similar results in terms of accuracy. An advanced Facial Emotion Recognition (FER) framework based on ResNet-50, the Convolutional Block Attention Module (CBAM), 3D Convolutional Neural Networks (3D CNN), and Ant Colony and Genetic Algorithm-based Target Optimization (AGTO) was evaluated on the KDEF dataset, achieving an accuracy of 98.35% [39]. The DCRNet framework achieved an accuracy of 96.25% on the KDEF dataset, demonstrating competitive performance compared with recent state-of-the-art facial expression recognition methods [40]. An innovative framework that combines data augmentation and super-resolution preprocessing with an ensemble of EfficientNetB0 and InceptionV3 backbones, followed by channel and spatial attention mechanisms and fully connected layers for emotion classification achieves an accuracy of 99.30% on the KDEF dataset [41]. Several lightweight transformer-based facial emotion recognition models, including Vision Transformer, Pooling-based Vision Transformer, Shifted Windows Transformer, Data-Efficient Image Transformer, and Cross-Attention Vision Transformer, were evaluated on the KDEF dataset. Among these, the Pooling-based Vision Transformer, which integrates convolutional and transformer components, achieved the highest accuracy of 0.9090, while the Shifted Windows Transformer obtained the lowest accuracy of 0.8434 [42].

In contrast, the proposed AU-ML framework demonstrates that interpretable handcrafted representations based on facial muscle movements can still achieve strong classification performance using comparatively lightweight machine learning methods. This may be particularly advantageous in real-time systems, embedded applications, or scenarios with limited computational capacity, where explainability and efficiency are important.

By exposing both the intermediate AU representations and the final prediction, this approach demonstrates transparency and interpretability absent from conventional end-to-end deep learning systems, whose internal decision-making processes often remain opaque. This implementation demonstrates that the proposed framework is not only computationally efficient but also suitable for deployment in practical affective computing applications requiring explainable emotion recognition.

Nevertheless, several limitations should be acknowledged. First, the KDEF dataset was collected under controlled laboratory conditions with posed emotional expressions, which may limit the generalizability of the results to spontaneous real-world emotional displays. Furthermore, the use of lateral images introduces additional challenges due to partial facial occlusion, potentially reducing the visibility

of some AUs. Although data augmentation increased dataset size and variability, augmented samples cannot fully replicate the diversity present in naturally occurring facial behavior. Additionally, the slight imbalance introduced by failed detections may have influenced classification outcomes to a limited extent.

Future research may focus on extending the proposed framework to more ecologically valid datasets containing spontaneous emotional expressions, varying illumination conditions, and natural head movements. Further investigation could also examine the integration of temporal information from video sequences, as dynamic facial changes may provide additional discriminative cues unavailable in static images. Another promising direction involves combining AU-based representations with deep learning embeddings or multimodal signals, such as vocal or physiological features, to improve robustness in complex real-world environments. Finally, additional optimization of feature selection strategies and hyperparameter tuning may further improve classification accuracy while maintaining computational efficiency.

Overall, the obtained results indicate that the proposed AU-ML framework constitutes an effective and computationally efficient approach for emotion recognition from facial expressions, with CatBoost emerging as the most suitable classification model among those evaluated.

5. Conclusion

This study proposed an AU-based machine learning (AU-ML) framework for emotion recognition from frontal facial expressions using the KDEF dataset. By leveraging OpenFace for AU extraction and applying manual feature selection, the original high-dimensional feature space was reduced to a compact and interpretable representation consisting of 34 features. Comparative evaluation of multiple machine learning algorithms demonstrated that ensemble methods achieved the strongest performance, with CatBoost emerging as the most effective classifier. The findings indicate that frontal facial expressions contain sufficient discriminative information for reliable emotion recognition, and that interpretable AU-based approaches can provide a computationally efficient alternative to resource-intensive deep learning methods. Furthermore, the development of a prototype application demonstrated the practical applicability and explainability of the proposed framework. Although the study was conducted under controlled laboratory conditions, the obtained results provide a strong foundation for future research on real-world and multimodal emotion recognition systems. Future work may involve assessing the proposed framework on more ecologically valid datasets and incorporating temporal information from video sequences to better capture dynamic facial expressions. In particular, further evaluation on spontaneous and in-the-wild FER datasets, such as CK+, AffectNet, RAF-DB, and FER2013, is recommended to better examine the generalizability of the approach. Furthermore, given its computational efficiency, the proposed approach could be further explored for real-time facial emotion recognition applications, particularly in low-resource or edge computing environments.

Author contributions: Conceptualization, methodology, NT and MT; validation, NT

and MT; formal analysis, MT and MM; investigation, NT and MM; data curation, NT; writing-original draft preparation, NT; writing-review and editing, MT; visualization, NT and MM. All authors have read and agreed to the published version of the manuscript.

Funding: This research and manuscript preparation were conducted without external funding or financial support.

Institutional review board statement: This study did not involve human participants or animal subjects and thus did not require ethical approval.

Informed consent statement: Not applicable.

Data availability statement: This study exclusively used data obtained from secondary sources through a comprehensive literature review. All referenced data are publicly accessible and have been appropriately cited within the manuscript.

Conflict of interest: The authors declare no conflict of interest.

AI use statement: While preparing this manuscript, the authors used Grammarly and ChatGPT solely for English language editing and polishing. All suggestions provided by these tools were carefully reviewed and revised by the authors, who take full responsibility for the accuracy, clarity, and integrity of the final content presented in this publication.

References

1. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. arXiv preprint. 2023. doi: 10.48550/arXiv.1706.03762
2. Deepika, Vashisth S, Sharma P. Effectiveness of Focal Loss for Traffic Sign Detection Using Deep Neural Networks. International Journal of convergence in healthcare. 2023; 3(2). doi: 10.55487/9rtyzm18
3. Xu Y, Lin YS, Zhou X, et al. Utilizing emotion recognition technology to enhance user experience in real-time. Computing and Artificial Intelligence. 2024; 2(1): 1388. doi: 10.59400/cai.v2i1.1388
4. Dhuheir M, Albaseer A, Baccour E, et al. Emotion Recognition for Healthcare Surveillance Systems Using Neural Networks: A Survey. arXiv preprint. 2021. doi: 10.48550/arXiv.2107.05989
5. Lima ES, Perico CP, Nichio BTL, et al. Emotional recognition technologies applied to health: Review and challenges. Brazilian Journal of Psychiatry. 2025; 47: e20243963. doi: 10.47626/1516-4446-2024-3963
6. Wu Y, Mi Q, Gao T. A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions. Biomimetics. 2025; 10(7): 418. doi: 10.3390/biomimetics10070418
7. Kalateh S, Estrada-Jimenez LA, Nikghadam-Hojjati S, et al. A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. IEEE Access. 2024; 12: 103976–104019. doi: 10.1109/ACCESS.2024.3430850
8. Kumar MJD, Rao MS, Narendra KC. Multimodal Emotion Recognition: A Comprehensive Survey of Datasets, Methods, and Applications. IEEE Access. 2025; 13: 201067–201097. doi: 10.1109/ACCESS.2025.3636186
9. Gautam C, Seeja KR. Facial emotion recognition using Handcrafted features and CNN. Procedia Computer Science. 2023; 218: 1295–1303. doi: 10.1016/j.procs.2023.01.108
10. Nithin DK, Sivakumar PB. Generic Feature Learning in Computer Vision. Procedia Computer Science. 2015; 58: 202–209. doi: 10.1016/j.procs.2015.08.054
11. Zhalgas A, Amirgaliyev B, Sovet A. Robust Face Recognition Under Challenging Conditions: A Comprehensive Review of Deep Learning Methods and Challenges. Applied Sciences. 2025; 15(17): 9390. doi: 10.3390/app15179390
12. Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001

IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 8–14 December 2001; Kauai, HI, USA. doi: 10.1109/CVPR.2001.990517
13. Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); 20–25 June 2005; San Diego, CA, USA. pp. 886–893. doi: 10.1109/CVPR.2005.177
14. Ali A, Nasir JA, Ahmed MM, et al. Machine Learning Based Statistical Analysis of Emotion Recognition using Facial Expression. RADS Journal of Biological Research & Applied Sciences. 2020; 11(1): 39–46. doi: 10.37962/jbas.v11i1.262
15. Ahonen T, Hadid A, Pietikäinen M. Face Recognition with Local Binary Patterns. In: Computer Vision - ECCV 2004, Lecture Notes in Computer Science. Springer; 2004. pp. 469–481. doi: 10.1007/978-3-540-24670-1_36
16. Asad M, Gilani SO, Jamil M. Emotion Detection through Facial Feature Recognition. International Journal of Multimedia and Ubiquitous Engineering. 2017; 12(11): 21–30. doi: 10.14257/ijmue.2017.12.11.03
17. Deepika, Vashisth S, Saurav S. Histogram of Oriented Gradients Based Reduced Feature for Traffic Sign Recognition. In: Proceedings of the 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI); 19–22 September 2018; Bangalore, India. pp. 2206–2212. doi: 10.1109/ICACCI.2018.8554624
18. Weng CH, Lai SH. Online facial expression recognition based on combining texture and geometric information. In: Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP); 27–30 October 2014; Paris, France. pp. 5976–5980. doi: 10.1109/ICIP.2014.7026206
19. Tico M, Haverinen T, Kuosmanen P. A Method of Color Histogram Creation for Image Retrieval. In: Proceedings of the NORSIG 2000 Nordic Signal Processing Symposium; 13–15 June 2000; Kolmården, Sweden. pp. 157–160. Available online: https://www.researchgate.net/publication/244449350_A_method_of_color_histogram_creation_for_image_retrieval
20. Lindeberg T. Orientation selectivity properties for the affine Gaussian derivative and the affine Gabor models for visual receptive fields. Journal of Computational Neuroscience. 2025; 53(1): 61–98. doi: 10.1007/s10827-024-00888-w
21. Sagonas C, Tzimiropoulos G, Zafeiriou S, et al. 300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge. In: Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops; 2–8 December 2013; Sydney, Australia. pp. 397–403. doi: 10.1109/ICCVW.2013.59
22. Liu C, Hirota K, Ma J, et al. Facial Expression Recognition Using Hybrid Features of Pixel and Geometry. IEEE Access. 2021; 9: 18876–18889. doi: 10.1109/ACCESS.2021.3054332
23. Haghpanah MA, Saeedizade E, Masouleh MT, et al. Real-Time Facial Expression Recognition using Facial Landmarks and Neural Networks. In: Proceedings of the 2022 International Conference on Machine Vision and Image Processing (MVIP); 23–24 February 2022; Ahvaz, Iran. pp. 1–7. doi: 10.1109/MVIP53647.2022.9738754
24. Kim KI, Kim J, Jung K. Recognition of facial images using support vector machines. In: Proceedings of the 11th IEEE Signal Processing Workshop on Statistical Signal Processing; 8 August 2001; Singapore. pp. 468–471. doi: 10.1109/SSP.2001.955324
25. Anggo M, Arapu L. Face Recognition Using Fisherface Method. Journal of Physics: Conference Series. 2018; 1028: 012119. doi: 10.1088/1742-6596/1028/1/012119
26. Kshirsagar VP, Baviskar MR, Gaikwad ME. Face recognition using Eigenfaces. In: Proceedings of the 2011 3rd International Conference on Computer Research and Development; 11–13 March 2011; Shanghai, China. pp. 302–306. doi: 10.1109/ICCRD.2011.5764137
27. Wagner U, Dürrschmid K, Pauser S. Emotion Recognition—Recent Advances and Applications in Consumer Behavior and Food Sciences with an Emphasis on Facial Expressions. In: Emotion Recognition—Recent Advances, New Perspectives and Applications. IntechOpen; 2023. doi: 10.5772/intechopen.110581
28. Shahzad HM, Bhatti SM, Jaffar A, et al. Hybrid Facial Emotion Recognition Using CNN-Based Features. Applied Sciences. 2023; 13(9): 5572. doi: 10.3390/app13095572
29. Kumar R, Corvisieri G, Fici TF, et al. Transfer Learning for Facial Expression Recognition. Information. 2025; 16(4): 320. doi: 10.3390/info16040320
30. Zhang K, Zhang Z, Li Z, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters. 2016; 23(10): 1499–1503. doi: 10.1109/LSP.2016.2603342
31. Garg D, Goel P, Pandya S, et al. A Deep Learning Approach for Face Detection using YOLO. In: Proceedings of the 2018 IEEE Punecon; 30 November–2 December 2018; Pune, India. pp. 1–4. doi: 10.1109/PUNECON.2018.8745376
32. Deng J, Guo J, Zhou Y, et al. RetinaFace: Single-stage Dense Face Localisation in the Wild. arXiv preprint. 2019.

doi: 10.48550/arXiv.1905.00641
33. Barrett LF, Lewis M, Haviland-Jones JM. Handbook of Emotions, 4th ed. Guilford Press; 2016.
34. Paul Ekman Group. Facial Action Coding System. Paul Ekman Group; n.d. Available online: https://www.paulekman.com/facial-action-coding-system/
35. KDEF & AKDEF. About KDEF. KDEF & AKDEF; n.d. Available online: https://kdef.se/home/aboutKDEF
36. Baltrusaitis T, Robinson P, Morency LP. OpenFace: An open source facial behavior analysis toolkit. In: Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV); 7–10 March 2016; Lake Placid, NY, USA. pp. 1–10. doi: 10.1109/WACV.2016.7477553
37. Baltrusaitis T, Robinson P, Morency LP. Constrained Local Neural Fields for Robust Facial Landmark Detection in the Wild. In: Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops; 2–8 December 2013; Sydney, Australia. pp. 354–361. doi: 10.1109/ICCVW.2013.54
38. CatBoost. CatBoost Is a High-Performance Open Source Library for Gradient Boosting on Decision Trees. CatBoost; n.d. Available online: httpscatboost.ai
39. Aly M, Alotaibi NS. A comprehensive deep learning framework for real time emotion detection in online learning using hybrid models. Scientific Reports. 2025; 15(1): 42012. doi: 10.1038/s41598-025-26381-7
40. Abdeldayem M, Badawy W, F. A. Hamed H, et al. Improving Facial Expression Recognition in real-world Environments. Statistics, Optimization & Information Computing. 2025; 14(6): 3546–3564. doi: 10.19139/soic-2310-5070-3171
41. Khan T, Yasir M, Choi C. Attention-enhanced optimized deep ensemble network for effective facial emotion recognition. Alexandria Engineering Journal. 2025; 119: 111–123. doi: 10.1016/j.aej.2025.01.078
42. Arslanoğlu MC, Acar H, Albayrak A. Face Expression Recognition via transformer-based classification models. Balkan Journal of Electrical and Computer Engineering. 2024; 12(3): 214–223. doi: 10.17694/bajece.1486140