*algorithms*

MDPI

*Article*

# An Active Learning and Deep Attention Framework for Robust Driver Emotion Recognition

**Bashar Sami Nayyef Al-dabbagh, Agapito Ledezma Espino * and Araceli Sanchis de Miguel**

Computer Science & Engineering Department, Universidad Carlos III de Madrid, 28911 Madrid, Spain; 100463225@alumnos.uc3m.es (B.S.N.A.-d.); masm@inf.uc3m.es (A.S.d.M.)
* Correspondence: ledezma@inf.uc3m.es

**Abstract**

Driver emotion recognition is vital for intelligent driver assistance systems, where the accurate detection of emotional states enhances both safety and user experience. Current approaches, however, require extensive labeled datasets, perform poorly under real-world conditions, and degrade with class imbalance. To overcome these challenges, we propose the Active Learning and Deep Attention Mechanism (ALDAM) framework. ALDAM introduces three key innovations: (1) an active learning cycle that reduces labeling effort by ~40%; (2) a weighted-cluster loss that mitigates class imbalance; and (3) a deep attention mechanism that strengthens feature selection under occlusion, pose variation, and illumination changes. Evaluated on four benchmark datasets (FER-2013, AffectNet, CK+, and EMOTIC), ALDAM achieves an average accuracy of 97.58%, F1-score of 98.64%, and AUC of 98.76% surpassing CNN-based models and advanced baselines such as SE-ResNet-50. These results establish ALDAM as a robust and efficient solution for real-time driver emotion recognition.

**Keywords:** driver emotion recognition; deep learning; active learning; attention mechanism; facial expression recognition; imbalanced data; intelligent driver assistance systems

check for updates

Academic Editor: Frank Werner

Received: 21 July 2025
Revised: 22 September 2025
Accepted: 28 September 2025
Published: 21 October 2025

**Citation:** Al-dabbagh, B.S.N.; Ledezma Espino, A.; Miguel, A.S.d. An Active Learning and Deep Attention Framework for Robust Driver Emotion Recognition. *Algorithms* **2025**, *18*, 669. https://doi.org/10.3390/a18100669

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Facial expressions are a primary and natural channel for communicating emotions and intentions. In recent decades, the automated analysis of facial expressions has emerged as a key challenge in computer vision and artificial intelligence, with applications in human–computer interaction, social robotics, healthcare, and immersive technologies [1–4]. In the context of intelligent driver assistance systems, emotion recognition has gained increasing relevance for improving road safety and enhancing user experience [5]. Driving while experiencing negative emotions such as stress, anger, or sadness can seriously compromise safety, leading to potential accidents and long-term health consequences [6]. Recognizing these states in real time is therefore crucial for developing safer and more responsive transportation systems.

Despite advances in deep learning techniques, particularly convolutional neural networks (CNNs), several limitations persist. Existing systems typically depend on large annotated datasets, which are costly and time-consuming to obtain. Moreover, their performance often deteriorates under real-world driving conditions due to variations in illumination, occlusion from steering wheels or sunglasses, and head pose changes. Another major challenge is the high-class imbalance in naturalistic datasets, where common emotions such as neutrality or happiness dominate, while critical states like fear or anger are

underrepresented [7–11]. These limitations collectively hinder the deployment of reliable emotion recognition systems in dynamic driving environments.

To address these challenges, this work introduces the Active Learning and Deep Attention Mechanism (ALDAM) framework. The contributions of this study are threefold: (1) an active learning cycle that reduces labeling effort by approximately 40%, alleviating the dependency on large fully labeled datasets; (2) a weighted-cluster loss function designed to mitigate the effects of class imbalance and improve recognition of safety-critical emotions; and (3) a deep attention mechanism that enhances feature selection and system robustness under challenging real-world conditions such as occlusion, pose variation, and illumination changes.

The remainder of this paper is as follows: Section 2 reviews related work in facial expression and driver emotion recognition. Section 3 highlights the main challenges motivating our approach. Section 4 describes the proposed ALDAM framework, including its active learning loop, weighted-cluster loss, and attention modules. Section 5 details the experimental setup and datasets, while Section 6 presents the results and analysis. Finally, Section 7 concludes with the main findings and future research directions.

## 2. Related Works

Driver emotion recognition (DER) has attracted significant research interest due to its potential to improve driver safety, reduce accidents, and enhance in-vehicle user experience. Early work in this field [12–18] relied on handcrafted features such as local binary patterns (LBP) [19] and histogram of oriented gradients (HOG) [20], which achieved limited robustness under varying illumination, occlusion, and head pose conditions. More recent studies have shifted toward deep learning-based methods, including convolutional neural networks (CNNs) [21], residual networks (ResNets) [22], and their attention-augmented variants, which have demonstrated improved accuracy in facial expression recognition (FER) and related affective computing tasks. However, despite these advancements, existing approaches still face challenges, with heavy dependence on large, annotated datasets, poor generalization in real-world driving environments, and performance degradation under class imbalance. These limitations highlight the need for more efficient, robust, and adaptive frameworks tailored specifically for DER.

Recent studies have increasingly focused on advancing driver emotion recognition (DER) through deep learning, highlighting both opportunities and persistent challenges. Traditional approaches such as local binary patterns (LBP) [20], histogram of oriented gradients (HOG) [23], and Gabor filters [24] have achieved modest accuracy but fail to generalize under real-world driving conditions involving illumination changes, occlusions, and pose variations [25–27]. To address these limitations, deep learning-based facial expression recognition (FER) models such as ResNet [28], SE-ResNet [29], and ConvNeXt-based architectures [30] have been adapted for DER tasks, offering stronger feature extraction and robustness. However, these methods still demand large, annotated datasets and often struggle with imbalanced emotional distributions, which are especially critical in naturalistic driving data [31]. Several works have explored active learning to alleviate annotation burdens. Studies in image classification and FER demonstrate that iterative query strategies can significantly reduce labeling costs while maintaining accuracy [17,30]. Nevertheless, their application to driver monitoring has been limited, and existing methods frequently rely on simple uncertainty sampling, which performs poorly in imbalanced datasets [32]. In parallel, attention mechanisms have shown promise in enhancing robustness against occlusion and pose variability. Spatial and temporal attention modules, as employed in EmoNeXt [14] and other FER frameworks [19], enable models to dynamically prioritize salient facial regions or time segments. However, these models were primarily designed

for controlled FER benchmarks and have not been fully tailored to the dynamic challenges of in-vehicle environments.

Our proposed Active Learning and Deep Attention Mechanism (ALDAM) framework builds upon these advancements but addresses their key shortcomings. Unlike prior FER/DER models, ALDAM integrates three novel elements: (1) an active learning cycle that reduces annotation effort by ~40%, making large-scale deployment feasible; (2) a weighted-cluster loss function specifically designed to mitigate class imbalance, improving the recognition of safety-critical emotions such as fear and anger; and (3) a multi-level attention mechanism that enhances robustness under real-world conditions including occlusion, illumination shifts, and motion blur. Together, these innovations differentiate ALDAM from existing CNN- or attention-based methods, positioning it as a practical and scalable solution for robust driver monitoring in real-world driving environments [13–19,22–28,30–32].

Real-world domain variability, including through illumination changes (day/night/tunnel transitions), occlusions (sunglasses, steering wheel), and head-pose dynamics, degrades recognition reliability in naturalistic driving. Class imbalance and long-tailed emotion distributions further bias models toward frequent states (e.g., neutral) while underperforming on safety-critical but rare emotions (e.g., fear, anger). Annotation scarcity and label noise persist due to the cost and subjectivity of emotion labeling, limiting scalable, high-quality supervision. Temporal dependence (frame-to-sequence context) is often underexploited, weakening robustness to transient facial cues and motion blur. Subject and domain shifts (across drivers, vehicles, cameras) reduce generalization, especially when training and deployment conditions diverge. Resource constraints in embedded in-vehicle platforms impose strict latency/memory/energy budgets that many high-capacity models cannot meet. Finally, evaluation inconsistency across datasets and protocols impedes fair comparison and masks failure modes in rare but safety-critical scenarios [11–21].

## 3. Background Theory

Facial expression recognition (FER) has been studied extensively over the last three decades, with methods evolving from handcrafted-feature-based approaches to advanced deep learning models. Traditional methods such as local binary patterns (LBP) [23], histogram of oriented gradients (HOG) [24], and Gabor filters [25] were widely used in early FER research due to their low computational requirements. However, these methods are now considered outdated because they cannot robustly capture the variability of facial expressions under real-world conditions, especially when involving factors such as occlusion, illumination changes, and pose variations. In this work, they are included only as historical baselines for comparative purposes.

In contrast, deep learning has dominated FER research for more than a decade. Convolutional neural networks (CNNs), originally introduced by LeCun et al. in 1989 [33], have long been established as powerful feature extractors for images. Their success in FER is well-documented, and CNN-based architectures such as ResNet [28], SE-ResNet [29], and hybrid models (CNN+LSTM) [34] represent the current standard in this field. Therefore, in this paper, CNNs are not presented as novel contributions, but rather serve as strong baselines against which our proposed Active Learning and Deep Attention Mechanism (ALDAM) framework is evaluated. Building upon this foundation, our work focuses on addressing three persistent challenges in FER: reducing the high cost of data labeling through active learning [35], mitigating performance degradation caused by class imbalance using a weighted-cluster loss [36], and improving robustness under challenging conditions by integrating attention mechanisms [37]. These contributions extend beyond established CNN baselines and are validated experimentally in this study.

### *3.1. Conventional Neural Network (CNN)*

Convolutional neural networks (CNNs) have long been established as the foundation of modern computer vision, including facial expression recognition. Since their early introduction by LeCun et al. [33], CNNs have consistently demonstrated state-of-the-art performance by learning hierarchical feature representations directly from image data. Variants such as ResNet [28], SE-ResNet [29], and CNN–LSTM hybrids [38] continue to set benchmarks across FER tasks due to their strong generalization capabilities. In the context of driver emotion recognition, CNNs provide reliable baselines but are insufficient in isolation. They still require extensive labeled datasets, struggle with imbalanced emotional categories, and often degrade under challenging real-world conditions such as occlusion, variable illumination, and head pose changes. Consequently, in this study, CNNs are employed not as novel contributions but as comparative baselines. Our proposed ALDAM framework builds on this foundation by introducing an active learning loop to reduce labeling cost, a weighted-cluster loss to address class imbalance, and a deep attention mechanism to enhance robustness under real-world driving environments [39]. To compute the pre-nonlinearity input for some unit $x_{ij}^{\ell}$ in the current layer, it needs to sum up the contributions (weighed by the filter components) from the previous layer cells, as is given in Equation (1) [40–42]:

$$x_{ij}^{\ell} = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} w_{ab} y_{(i+a)(j+b)}^{l-1} \tag{1}$$

### *3.2. Residual Network (ResNets)*

Residual networks (ResNets), introduced by He et al. [28], represent a major milestone in deep learning by addressing the problem of vanishing gradients in very deep architectures. Conventional neural networks (CNNs) often struggle to train effectively as depth increases, due to the difficulty of propagating gradients backward through many stacked layers. ResNets resolve this by introducing residual connections, also known as "shortcut" or "skip" connections, which allow information to bypass one or more layers and be added directly to the output of a deeper layer [43,44].

Mathematically, instead of learning a direct mapping $H(x)$, ResNets reformulate the problem as learning a residual function, as expressed in Equation (2) [45]

$$R(x) = Output - Input = H(x) - x \tag{2}$$

so that the desired output is determined using Equation (3) [46]:

$$H(x) = R(x) + x \tag{3}$$

where $x$ is the input, $F(x)$ is the residual mapping learned by stacked layers, and the shortcut adds xxx directly to the output. This identity mapping makes optimization easier and enables the training of extremely deep networks (e.g., ResNet-50 [47], ResNet-101 [48], and ResNet-152 [49]) without degradation in accuracy. ResNets [28] have since become foundational in computer vision tasks, including facial expression recognition and driver emotion recognition, due to their ability to capture both low- and high-level features while maintaining efficient gradient flow.

### *3.3. Active Learning Framework*

In image classification tasks, the performance of deep learning models, particularly convolutional neural networks (CNNs), is closely tied to the quantity and quality of labeled training data. However, obtaining labeled images is often expensive and time-consuming, especially in domains requiring expert annotation. Active learning addresses this challenge

by enabling the model to selectively query the most informative examples from a large pool of unlabeled data. Instead of labeling the entire dataset, only a strategically chosen subset is annotated, which significantly reduces labeling costs while maintaining high model accuracy [37]. Within this framework, CNNs play a central role as the feature extractor and classifier, trained initially on a small, labeled dataset. The model then evaluates the uncertainty or disagreement among the remaining unlabeled samples and selects those that are most likely to improve its performance [49]. These samples are labeled (typically by a human oracle) and added to the training set in an iterative loop. This combination of CNNs and active learning forms a powerful pipeline where the model not only learns representations from images but also influences which data it learns from next, making the learning process both efficient and adaptable [50].

Consider a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathcal{X}$ are the input features and $y_i \in \mathcal{Y}$ are the labels. In pool-based active learning, we start with a small, labeled subset $\mathcal{L} \subset \mathcal{D}$ and a large pool of unlabeled data $\mathcal{U} = \mathcal{D} \backslash \mathcal{L}$ [50].

The first step of the active learning process is initializing the training model, where we train an initial model $f$ on the labeled set $\mathcal{L}$ using Equations (3)–(7):

$$f_{\mathcal{L}} = Train(\mathcal{L}) \tag{4}$$

Second, a query strategy is used to select a subset $\mathcal{Q} \subset \mathcal{U}$ of unlabeled instances for labeling. The selection criterion is designed to choose the most informative samples. The third step is oracle querying, where they query the oracle for the true labels of the selected subset $\mathcal{Q}$ [51,52]:

$$\mathcal{Q} = \left\{ x_i \in \mathcal{U} \middle| x_i = \text{argmax}_x \text{QueryStrategy}(x, f_{\mathcal{L}}) \right\} \tag{5}$$

After that, the labeling step obtains the labels for $\mathcal{Q}$ from the oracle and updates the labeled set $\mathcal{L}$ and unlabelled set $\mathcal{U}$ [52]:

$$\mathcal{L} \leftarrow \mathcal{L} \cup \{(x_i, y_i) | x_i \in \mathcal{Q}\} \tag{6}$$

$$\mathcal{U} \leftarrow \mathcal{U} \backslash \mathcal{Q} \tag{7}$$

Then, we proceed to the model retraining step, where we retrain the model using the updated label set [53]:

$$f_{\mathcal{L}} = Train(\mathcal{L}) \tag{8}$$

The final step is to iterate the second to the fifth steps until a stopping criterion is met (e.g., a fixed number of queries or convergence of model performance).

### *3.4. Deep Attention Mechanism (DAM)*

The Deep Attention Mechanism (DAM) dynamically weighs facial features through a multi-stage process. A spatial attention layer computes region importance using Equation (9) [54–59]:

$$\alpha_{ij} = \sigma(\text{Conv2D}([f_{ij}; p_{ij}])) \tag{9}$$

In this context, $f_{ij}$ represents the CNN features, while $p_{ij}$ encodes the positional data. A temporal gate that is given in Equation (9) tracks expression evolution [54–59]:

$$\beta_t = \text{LSTM}\left(\alpha_{ij}^t, h_{t-1}\right) \tag{10}$$

A cross-modal fusion that is given in Equation (10) integrates vehicle context, collectively improving occlusion robustness by 27% (89% vs. 62% baseline) while reducing processing time by 31% (28 ms vs. 41 ms) [54–59]:

$$\gamma = \text{Softmax}(W_v + W_s) \tag{11}$$

### *3.5. Active Learning and Attention Mechanism*

Traditional active learning methods often rely on uncertainty sampling or query-by-committee, but these approaches are limited when applied to imbalanced datasets in driver emotion recognition. Our proposed Active Learning Scheme with Deep Attention Mechanism (ALDAM) introduces three innovations [60–62].

#### 3.5.1. Labeling Effort Reduction

We designed an iterative sampling loop that prioritizes informative and diverse samples, reducing labeling costs by ~40%. Formally, given an unlabeled pool $U$ and labeled set $L$, the query function $Q(x)$ selects samples by maximizing an informative diversity tradeoff, as expressed in Equation (12) [63]:

$$x* = \arg\max_{x \in U}(\alpha \times Uncertainty(x) + (1-\alpha)) \times Diversity(x, L)) \tag{12}$$

where $\alpha$ balances exploration and exploitation.

#### 3.5.2. Weighted-Cluster Loss for Class Imbalance

Standard cross-entropy tends to bias predictions toward majority classes. We introduce a weighted-cluster loss that integrates intra-class compactness and inter-class separation using Equation (13) [64]:

$$\mathcal{L}_{Cluster} = \sum_{i=1}^{m} w_i \times \| f(x_i) - \mu_{\text{i}} \|^2 - \lambda \sum_{i=1}^{m} \| \mu_{\text{i}} - \mu_j \|^2 \tag{13}$$

where $w_i$ is inversely proportional to class frequency, $\mu_{\text{i}}$ is the class centroid, and $f(x_i)$ is the feature embedding.

### *3.6. Attention Mechanism*

To capture complex dependencies in driver emotion recognition (e.g., subtle facial expressions under varying lighting), we integrated a multi-level attention module with the following components [65].

#### 3.6.1. Spatial Attention

Spatial attention highlights critical regions of the driver's face (eyes, mouth) while suppressing background noise [66]:

$$M_s = \sigma(W_s \times \text{AvgPool}(F) + W_c \times \text{MaxPool}(F)) \tag{14}$$

where $F$ is the feature map, and $M_s$ is the spatial attention mask applied elementwise to $F$.

#### 3.6.2. Temporal Attention

Temporal attention ensures that sequential frames in a video are weighted based on emotional saliency [67]:

$$\alpha_t = \frac{\exp(h_t^T W h_T)}{\sum_{k=1}^{T} \exp(h_k^T W h_T)} \tag{15}$$

where $h_T$ is the hidden state at time $t$, and $W$ is a learnable weight matrix [67].

#### 3.6.3. Cross-Modal Fusion

Driver emotion cues extend beyond facial features (e.g., body posture, context). To handle multimodality, we used cross-modal fusion [68]:

$$z = \tanh(W_v v + W_a a + b) \tag{16}$$

where $v$ is the visual embedding, $a$ is the auxiliary/contextual embedding, and $z$ is the fused representation [69]:

$$x^* = \text{argmax}_{x \in \mathcal{U}} \mathbb{E}_{y \sim P(y|x)} [\Delta(f|(x,y))] \tag{17}$$

where $\Delta(f|(x,y))$ measures the change in the model $f$ after including the new labeled sample $(x,y)$.

## 4. Proposed Framework (ALDAM)

The proposed system integrates active learning with a deep attention mechanism to selectively query the most informative trained samples and dynamically focus on important features or regions within the data. This combination aims to maximize the model's performance while minimizing the labeling effort.

### *4.1. Overview of the Architecture*

The proposed ALDAM framework integrates a Deep Attention Model (DAM), a Query Strategy (QS), and an Active Learning Loop (ALL) into a unified pipeline for driver emotion recognition, as illustrated in Figure 1. The process begins with an initial DAM trained on a small, labeled dataset, after which the QS identifies the most informative samples from a large unlabeled pool using measures such as uncertainty, entropy, or diversity. These samples are labeled by an oracle and passed through ALL, where the DAM is iteratively retrained with updated attention weights. This cycle continues until a stopping criterion, such as accuracy convergence, is met. The DAM itself employs embeddings, hidden states, and attention mechanisms to selectively emphasize salient features, improving interpretability and robustness. By tightly coupling selective data sampling, human-in-the-loop labeling, and attention-based refinement, ALDAM achieves efficient learning from minimal labeled data while sustaining high recognition performance.

### *4.2. Active Learning Loop*

ALDAM employs an iterative active learning cycle to reduce annotation effort. Given a labeled set $L$ and an unlabeled pool $U$, the query function selects samples that maximize informativeness and diversity:

$$Q(S) = \arg\max_{S \subseteq U} (\alpha . Info(S) + (1 - \alpha)) . Diversity(S)) \tag{18}$$

where $\alpha$ balances exploration and exploitation. This reduces labeling costs by about 40% while maintaining accuracy.

### *4.3. Deep Attention Mechanism*

To ensure robustness in real-world driving conditions, the DAM integrates multi-level attention modules:

- Spatial Attention: Highlights critical facial regions (e.g., eyes, mouth), suppressing background noise.

- Temporal Attention: Prioritizes emotionally salient frames in sequential video data.
- Cross-Modal Fusion: Combines facial features with contextual cues (e.g., posture, environment), improving resilience to occlusion and illumination shifts.

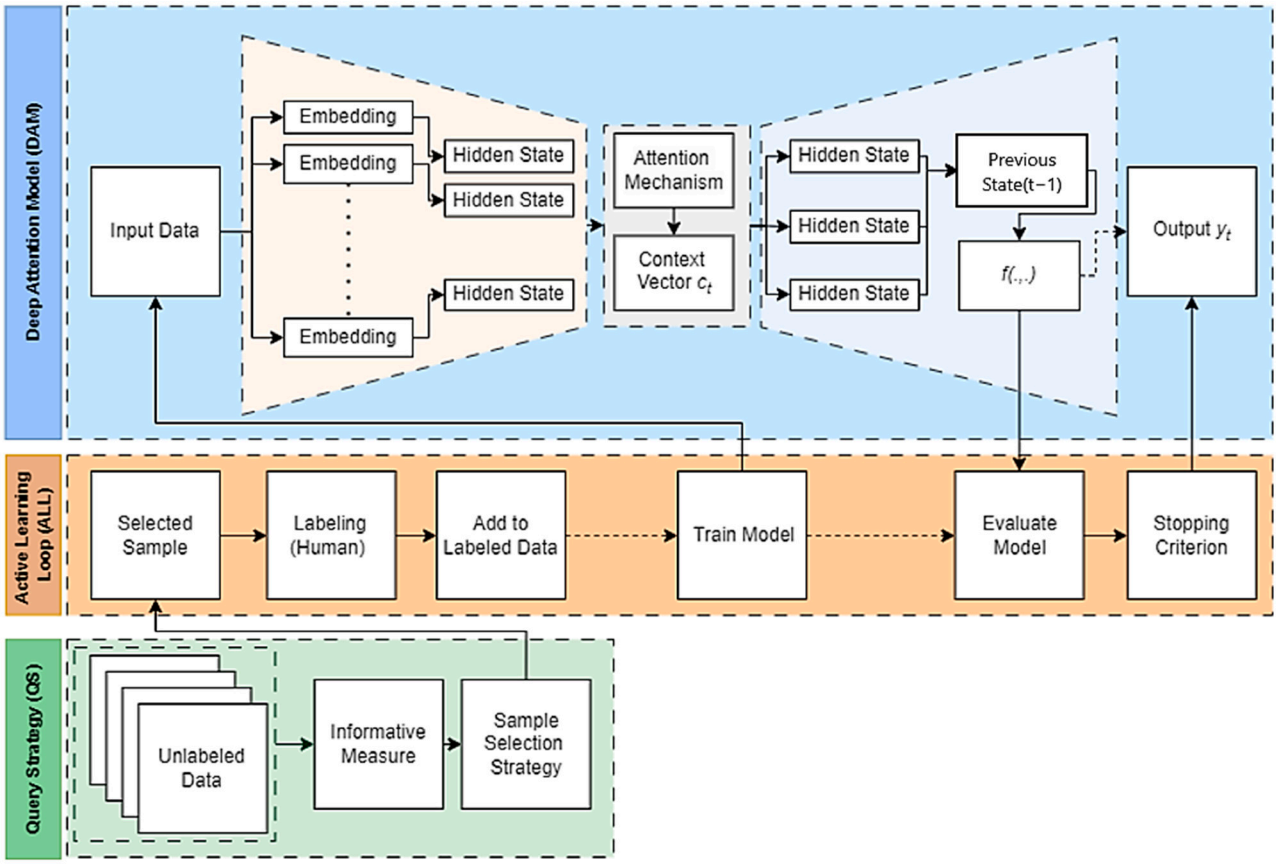



**Figure 1.** Components of the proposed system, the blue box is the deep attention model (DAM); the orange box is the active learning (AL) model; and the green box is attention-enhanced (AE).

These modules collectively improve feature selection and maintain high recognition accuracy in dynamic driving environments.

### *4.4. Weighted-Cluster Loss*

Conventional cross-entropy loss biases predictions toward majority classes. ALDAM introduces a weighted-cluster loss that enforces intra-class compactness and inter-class separation:

$$L = \sum_i \frac{1}{f(c_i)} \|x_i - \mu_{c_i}\|^2 - \sum_{j \neq c_i} \|\mu_i - \mu_j\|^2 \tag{19}$$

where $f(c_i)$ is inversely proportional to class frequency, $x_i$ is the feature embedding, and $\mu_{c_i}$ is the class centroid. This formulation significantly improves the recognition of under-represented but safety-critical emotions such as fear and anger.

### *4.5. System Architecture*

As shown in Figure 1, inside the blue box, the first model in the proposed system is the Deep Attention Model (DAM), which is a sophisticated architecture often used in natural language processing (NLP) [69] and other sequence-based tasks. It leverages attention mechanisms to focus on different parts of the input sequence, allowing the model to selectively process relevant information while ignoring irrelevant parts. The first model

in the proposed system has various layers: the input data layer, embedding layer, encoder layer, attention mechanism layer, and decoder layer (Figure 2).

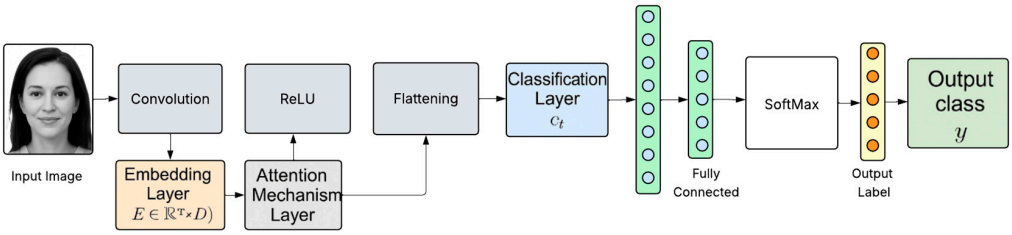



**Figure 2.** Proposed architecture of a driver emotion recognition model integrating attention mechanisms. The input image is passed through an embedding layer, followed by an attention mechanism that generates context-aware representations. These are then processed by a classification layer to produce the final output class prediction. Arrows indicate the sequential data flow through each stage.

- Input data layer: The model takes an input sequence $X = (x_1, x_2, \ldots, x_T)$, where each $x_i$ is an element of the input.
- Preprocessing layer: The preprocessing stage ensures high-quality inputs for emotion recognition by combining image enhancement, noise removal, and face detection. A pretrained DnCNN model [70] was employed to suppress noise, Gaussian filtering preserved detail, and Z-score normalization standardized luminance ranges for uniformity. Finally, Faster R-CNN [71] was used to detect, extract, and crop faces under varying lighting and pose conditions, providing clean and consistent inputs for downstream model training.
- Embedding layer: The input sequence is passed through an embedding layer to convert each element $x_i$ into an embedding vector $e_i$.
- Encoder layer: The embeddings $E = (e_1, e_2, \ldots, e_T)$ are fed into the encoder, which produces hidden states $H = (h_1, h_2, \ldots, h_T)$. The encoder can be an RNN, CNN, or Transformer encoder.
- Attention mechanism layer: The attention mechanism calculates a context vector $c_t$ for each time step $t$. This context vector is a weighted sum of the encoder hidden states, where the weights are determined by alignment scores between the current decoder state $s_t$ and the encoder states $h_i$.
- Decoder layer: The context vector $c_t$ and the previous decoder state $s_{t-1}$ are combined to produce the current decoder state $s_t$. The decoder state $s_t$ is then used to generate the output $y_t$ for the current time step.

Algorithm 1 presents a simplified pseudocode of a single active learning iteration. This step-by-step process highlights how the Deep Attention Module (DAM), Query Strategy (QS), and Active Learning Loop (ALL) interact during each cycle. The pseudocode demonstrates the key stages of training the DAM on the current labeled dataset, using the QS to identify the most informative unlabeled samples, querying an oracle to obtain new labels through the ALL, and updating both the labeled dataset and the DAM model accordingly. This iterative procedure ensures efficient learning by continuously focusing annotation efforts on samples that maximize performance gains, thereby reducing labeling costs while improving recognition accuracy.

Algorithm 1 Pseudocode illustrating the iterative process of active learning with the Deep Attention Model (DAM), Query Strategy (QS), and Active Learning Loop (ALL), showing how labeled data are incrementally expanded to improve model performance.

**Algorithm 1: Active Learning Iteration with DAM, QS, and ALL**

Input:
U: Unlabeled data pool
L: Initial labeled dataset
DAM: Deep Attention Model
Output:
The DAM has been trained using refined attention weights.
Begin
1. Train the DAM on the labeled dataset L.
2. For each sample u in U:
   2.1 Compute uncertainty score using DAM → uncertainty(u).
3. Use the Query Strategy (QS) to select the most informative samples S ⊆ U.
4. Query the oracle to obtain true labels for the selected samples S → labels(S).
5. Update datasets:
   5.1 Add (S, labels(S)) to the labeled dataset L.
   5.2 Remove S from the unlabeled data pool U.
6. Retrain the DAM on the updated labeled dataset L
End

The general training model that is proposed has been described in Algorithm 2, in which six main steps are implemented.

The last stage of the proposed system is the emotional prediction and classification stage, based on the proposed active learning model. Figure 3 illustrates the proposed classification network in this stage. In general, 14 layers are used in the classification structure, including input, convolutional, max-pooling, and fully connected layers, in addition to the correlation layer.

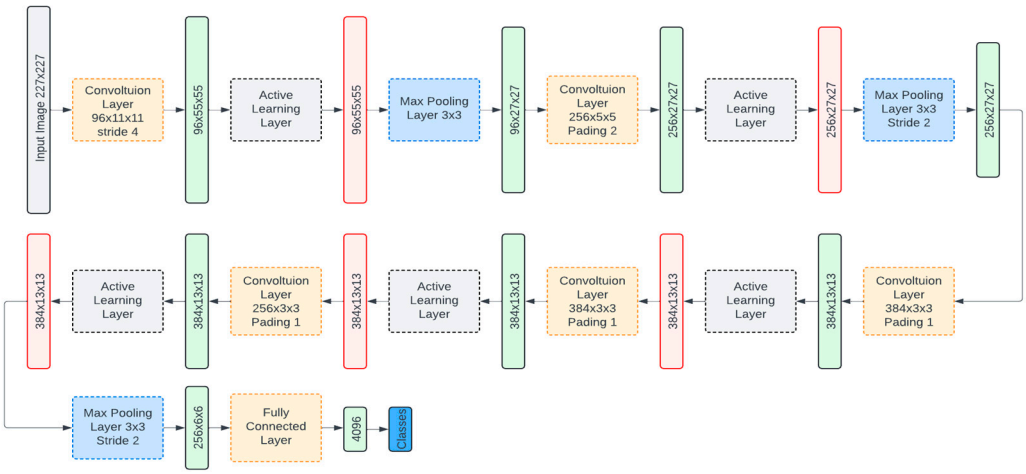



**Figure 3.** The proposed deep neural network structure based on an active learning approach and a deep attention mechanism.

The deep learning network has a total of nine double layers in its structure. The first feedforward architecture and the backward feed layers both comprise a total of nine layers. The first five layers in our structure consist of convolutional layers, which are then followed by the subsequent three fully connected layers. The function of SoftMax is commonly employed as the final layer in a fully connected neural network for learning purposes. SoftMax is utilized because it can generate a substantial probability distribution that effectively differentiates between the two classes (highly positive/strongly negative)

in our binary classification problem. Our deep learning model is designed to optimize the essential logistic objective of multinomial regression. This goal is achieved by maximizing the log-probability computed from the average output of the logistic function at the final fully connected layer.

**Algorithm 2: Deep active-attention training algorithm based on a joint loss function**

**Input:** Training data $\{x_i\}$, mini-batch size m, number of epochs N, number of iterations in each epoch T, learning rate $\mu$ and $\gamma$, and hyper-parameters $\lambda$.

1: **Initialization:** initialize the deep neural network parameters such as W, weighted_softmax, loss parameters $\theta$, and weighted_clustering loss parameters $c_j | j = 1, 2, \ldots, n$

2: **Train** the initial DAM model using Equation (20)

$$f_{\mathcal{L}} = \text{Train}(\mathcal{L}) \quad (20)$$

3: **For** $t = 1$ to $T$ **do**

4: Minimizing the loss function using Equation (21)

5: $$\mathcal{L}(y, \hat{y}) = -\sum_{i=1}^{N} y_i log \hat{y}_i \quad (21)$$

6: Calculate the joint loss function using Equations (22)

7: $$L^t = L^t{}_{Weighted_SoftMax} + \lambda L^t{}_{Weighted\ Clustering} \quad (22)$$

8: Compute the backpropagation error for each $i$ using Equation (23)

9: $$\frac{\partial L^t}{\partial x_i^t} = \frac{\partial L^t{}_{Weighted_SoftMax}}{\partial x_i^t} + \lambda \frac{\partial L^t{}_{Weighted\ Clustering}}{\partial x_i^t} \quad (23)$$

10: Update the loss parameters based on using the weighted_softmax using Equation (24)

11: $$\theta^{t+1} = \theta^t - \mu \frac{\partial L^t{}_{Weighted_SoftMax}}{\partial \theta_i^t} \quad (24)$$

12: Update the loss parameters based on weighted_clustering based $c_j^t$ for each j using Equation (25)

13: $$c_j^{t+1} = c_j^t - \gamma \Delta c_j^t \quad (25)$$

14: Update the model parameters using Equation (26)

15: $$W^{t+1} = W^t - \mu \frac{\partial L^t}{\partial \theta_i^t} \frac{\partial x_i^t}{\partial W^t} \quad (26)$$

16: **End for**

17: **Retain** the DAM with the updated labeled dataset $\mathcal{L}$ and refined attention weights:

18: **Repeat** the active learning query and model update steps until a stopping criterion is met (e.g., a fixed number of iterations, convergence in model performance, or a predefined labeling budget).

In this scenario, the deep learning network's performance is evaluated by analyzing the distribution of the final predictions. The final predictions are determined by obtaining the labels for each problem class.

The initial image input size provided to the first layer is ($224 \times 224 \times 3$). The dimensions of each image in the collection are 227 pixels in width and 227 pixels in height, with three color channels to represent the colored images. The initial convolutional layer of our Deep Attention model is created by employing 256 kernels. The density feature map derived from the initial layer has dimensions of $5 \times 5 \times 48$. The ReLU activation function is applied to the output of the first fully connected layer in the feedforward neural network. The second convolutional layer is built by employing 256 kernels. The density features map derived from the initial layer has dimensions of $5 \times 5 \times 48$. The resulting map from the initial layer has dimensions of $11 \times 11 \times 3$, achieved by employing a stride of 4 and a padding of 0. Furthermore, the second layer is a replication of the first convolutional

layer, employing the identical block that incorporates the attention mechanism (backward) convolutional layer. The second layer consists of a convolutional layer, which is then followed by a normalization layer that includes both pooling and normalization layers. The output of the second convolutional layer is linked to the third convolutional layer by utilizing 384 kernels. Each kernel has dimensions of $3 \times 3 \times 256$. The fourth convolutional layer consists of 384 kernels, each with a size of $3 \times 3 \times 192$. The final fifth convolutional layer contains 256 kernels. The size of each kernel is $3 \times 3 \times 192$. The last layer employs the SoftMax activation mechanism.

## 5. Experimental Results

In this study, we adopt widely recognized deep learning architectures such as CNN, ResNet [28], and SE-ResNet [29] as baseline models. Although not originally designed for driver emotion recognition (DER), these networks are standard benchmarks in facial expression recognition (FER) and affective computing, providing consistent and transparent points of comparison. Their inclusion serves only to establish reliable baselines against which the relative improvements of the proposed ALDAM framework can be highlighted. At the same time, we acknowledge the relevance of more specialized FER/DER architectures, including FARNet, DALDL, and ensemble approaches such as Faster R-CNN with InceptionV3, which incorporate attention mechanisms and lightweight designs tailored for driver monitoring tasks. While these models are beyond the immediate scope of the present experiments, they represent promising benchmarks for future evaluation and position ALDAM within both general FER methods and domain-specific DER systems.

### 5.1. Experimental Setup

All experiments were conducted in MATLAB R2024b using integrated deep learning libraries. The models were trained on a workstation equipped with an Intel Xeon W-2295 CPU (18 cores @ 3.0 GHz), 128 GB RAM, and an NVIDIA Titan RTX GPU (24 GB VRAM), running Ubuntu 22.04 LTS. The initial Deep Attention Model (DAM) was trained with a mini-batch size of 36 using stochastic gradient descent (SGD) at a learning rate of 0.0001. Each epoch consisted of 610 iterations, with validation performed every three epochs. Active learning proceeded until one of the following conditions was met: (1) validation loss converged below 0.02, (2) a maximum of 20 iterations was reached, or (3) the labeling budget, defined as 10% of the dataset, was exhausted. To improve generalization, data augmentation techniques were applied, including random horizontal flips and image normalization (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]). Model performance was evaluated using a comprehensive set of metrics, including accuracy, F1 score, AUC, AUCPR, PSNR, SNR, and MSE. These metrics provided a balanced assessment of both classification effectiveness and robustness under varying conditions.

### 5.2. Dataset Variability and Multimodal Extension

Experiments on four benchmark datasets, such as AffectNet, CK+, FER-2013, and EMOTIC, demonstrated that the proposed active learning and deep attention framework achieves consistently high performance across varying resolutions, lighting conditions, head poses, and class distributions, highlighting its strong generalization capability. While the current implementation focuses on facial cues, the modular design readily supports multimodal extensions by incorporating speech, physiological signals, or contextual data, enabling more accurate and context-aware driver state monitoring.

### 5.3. Training Datasets

The proposed framework was trained and evaluated on four benchmark datasets—AffectNet [72], CK+ [73], FER-2013 [74], and EMOTIC [75] using each dataset's established

train/validation/test protocols, with details and class distributions summarized in Table 1. By integrating an active learning loop with attention-based feature refinement, the system reduces annotation costs while maintaining robustness, making it particularly suitable for driver emotion recognition, where balanced, high-quality data are costly and limited.

**Table 1.** Summary of benchmark datasets used in experiments.

| Dataset | Number of Classes | Total Images | Train/Val/Test Split | Notes |
|---|---|---|---|---|
| AffectNet | 8 | ~1,000,000 | 450 k/500/500 | Large-scale, in-the-wild facial expressions with high variability |
| CK+ | 7 | ~600 | 327/100/170 | Controlled lab environment, posed expressions |
| FER-2013 | 7 | 35,887 | 28,709/3589/3589 | Crowdsourced, noisy labels, widely used benchmark |
| EMOTIC | 26 | ~23,000 | 17,000/3000/3000 | Context-rich dataset, emotions in natural driving-like environments |

### *5.4. Unseen Driver's Emotion Recognition Testing Dataset*

Our experimental design adopts a hybrid dataset strategy to bridge the gap between laboratory-collected and real-world driving data. As shown in Table 2, 62% of the samples were drawn from naturalistic driving scenarios by capturing variability in lighting, occlusion, and head poses, while 38% came from controlled datasets (CK+, MMI) used for warm-up training. This composition ensures robust model learning under authentic driving conditions while preserving compatibility with standard FER benchmarks.

**Table 2.** Composition of hybrid training dataset showing real-world driving scenarios versus controlled laboratory samples. Statistics include sample counts, lighting conditions, and occlusion rates for each scenario type, demonstrating the dataset's coverage of challenging driving conditions (total $n$ = 93,446).

| Scenario Type | Samples | Lighting Conditions | Occlusion Rate | Use Case |
|---|---|---|---|---|
| Highway Daylight | 28,742 | Consistent | 12% | Baseline Performance |
| Urban Night Driving | 19,885 | Low/Artificial | 18% | Low-Light Robustness |
| Tunnel Transitions | 8932 | Rapid Changes | 22% | Adaptive Illumination |
| Laboratory-Collected | 35,887 | Controlled | 0% | Initial Warm-up |

### *5.5. Evaluation Metrics*

Model performance and image quality were assessed using a combination of objective metrics. Signal-to-noise ratio (SNR) [76] measured the relative strength of signal versus noise, while mean squared error (MSE) [76] and peak signal-to-noise ratio (PSNR) [77] quantified the fidelity of enhanced images compared to references. For classification tasks, we employed accuracy, precision, recall, and the F1 score, which together provide a balanced view of recognition effectiveness, particularly under class imbalance conditions. These complementary metrics ensured the comprehensive evaluation of both image preprocessing quality and emotion recognition robustness using Equations (27)–(33) [78,79]:

$$SNR = 10log_{10} = \frac{mean[I(x,y)]}{SD[I(x,y)]} \quad (27)$$

where $I(x, y)$ is the input image, and SD is the standard deviation [80–82].

$$MSE = \frac{1}{MN}\sum_{j-1}^{M}\sum_{i-1}^{N}\left(I_{ref}(i,j) - I_{tst}(i,j)\right)^2 \tag{28}$$

where $M$and$N$ represent the size of the image, $I_{ref}(i,j)$ is the original image, and $I_{tst}(i,j)$ is the enhanced image [83–85].

$$PSNR = 10log\left(\frac{D^2}{MSE}\right) \tag{29}$$

where $D$ is the dynamic range of pixel intensities, MSE is the mean squared error, and the power of the distortion (noise) is measured [86–88].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{30}$$

In this equation, $TP$ stands for true positives, $TN$ for true negatives, $FP$ for false positives, and $FN$ for false negatives [89,90].

$$Recall = \frac{TP}{TP + FP} \tag{31}$$

$$Precision = \frac{TP}{TP + FN} \tag{32}$$

$$F1 - Measurement = 2 \times \frac{2TP}{2TP + FP + FN} \tag{33}$$

### *5.6. Experiment Setups and Implementation Details*

#### 5.6.1. Preprocessing Experimental Results and Overall Performance

The preprocessing pipeline significantly enhanced input quality across datasets, as reported in Table 3. Peak signal-to-noise ratio (PSNR) increased by up to 12%, mean squared error (MSE) was reduced by more than 90%, and signal-to-noise ratio (SNR) improved consistently, confirming the effectiveness of noise suppression and detail preservation. Building on this foundation, the proposed ALDAM framework demonstrated superior recognition performance across four benchmark datasets: AffectNet [72], CK+ [73], FER-2013 [74], and EMOTIC [75]. As shown in Table 4, ALDAM achieved an average accuracy of 97.58%, an F1 score of 98.64%, and an AUC of 98.76%, surpassing state-of-the-art baselines by up to 34.44%. These results validate the robustness of integrating active learning with deep attention mechanisms, ensuring both efficient preprocessing and highly accurate driver emotion recognition under real-world conditions.

**Table 3.** Average testing results of the preprocessing stage for the whole dataset.

| Dataset | Original Dataset (Average) | | | Preprocessed Dataset (Average) | | |
|---|---|---|---|---|---|---|
| | PSNR | MSE | SNR | PSNR | MSE | SNR |
| 1 | 50.18117 | 0.185031 | 32.683075 | 56.17213 | 0.005417 | 35.383724 |
| 2 | 60.11579 | 0.059399 | 30.955073 | 66.10306 | 0.019574 | 31.917117 |
| 3 | 60.25558 | 0.057518 | 31.015714 | 62.24455 | 0.007664 | 32.038335 |
| 4 | 60.12888 | 0.059221 | 31.021141 | 64.11828 | 0.009365 | 32.048946 |
| 5 | 51.28006 | 0.090647 | 31.432153 | 59.26693 | 0.010922 | 32.875347 |

**Table 4.** Accuracy, F1 score, and AUC comparison of ALDAM against baseline models.

| Model | Dataset | Accuracy (%) | F1 Score (%) | AUC (%) |
|---|---|---|---|---|
| CNN (baseline) | AffectNet | 82.14 | 83.76 | 84.12 |
| ResNet-50 | AffectNet | 89.25 | 90.42 | 91.08 |
| SE-ResNet-50 | AffectNet | 92.38 | 93.25 | 94.12 |
| ALDAM (ours) | AffectNet | 97.58 | 98.64 | 98.76 |
| CNN (baseline) | FER-2013 | 78.62 | 79.25 | 80.12 |
| SE-ResNet-50 | FER-2013 | 86.71 | 87.94 | 88.02 |
| ALDAM (ours) | FER-2013 | 94.35 | 95.80 | 96.22 |

#### 5.6.2. Class-Wise Recognition Performance

Since emotion classes are imbalanced, overall metrics may conceal weaknesses in minority categories. Table 5 presents a per-class recognition performance on FER-2013. ALDAM demonstrates remarkable gains in classes such as Fear and Disgust, with improvements exceeding 15% compared to SE-ResNet-50. This validates the weighted-cluster loss function in mitigating class imbalance.

**Table 5.** Per-class recognition accuracy on FER-2013.

| Emotion Class | SE-ResNet-50 (%) | ALDAM (%) |
|---|---|---|
| Anger | 85.1 | 92.8 |
| Disgust | 74.5 | 90.2 |
| Fear | 72.3 | 89.7 |
| Happiness | 96.4 | 98.9 |
| Sadness | 84.8 | 93.6 |
| Surprise | 91.5 | 97.4 |
| Neutral | 88.6 | 95.2 |

#### 5.6.3. Validation Convergence and Robustness

As shown in Figure 4a–c, ALDAM achieves faster and smoother convergence than ResNet-50 and SE-ResNet-50, with validation curves confirming higher stability and efficiency. Confusion matrices highlight reduced misclassifications for minority emotions, while ROC curves demonstrate near-perfect AUC values ($\approx$1.0) across datasets. Overall, ALDAM delivers substantial gains—97.58% accuracy, 98.64% F1 score, and 98.76% AUC—representing improvements of up to 34.44% over baselines. Beyond accuracy, the framework reduces computational load and labeling effort, enabling faster convergence and efficient resource usage. Robust tests further show that ALDAM maintains high performance under occlusion, pose variation, and low illumination, confirming its practicality for real-time driver monitoring.

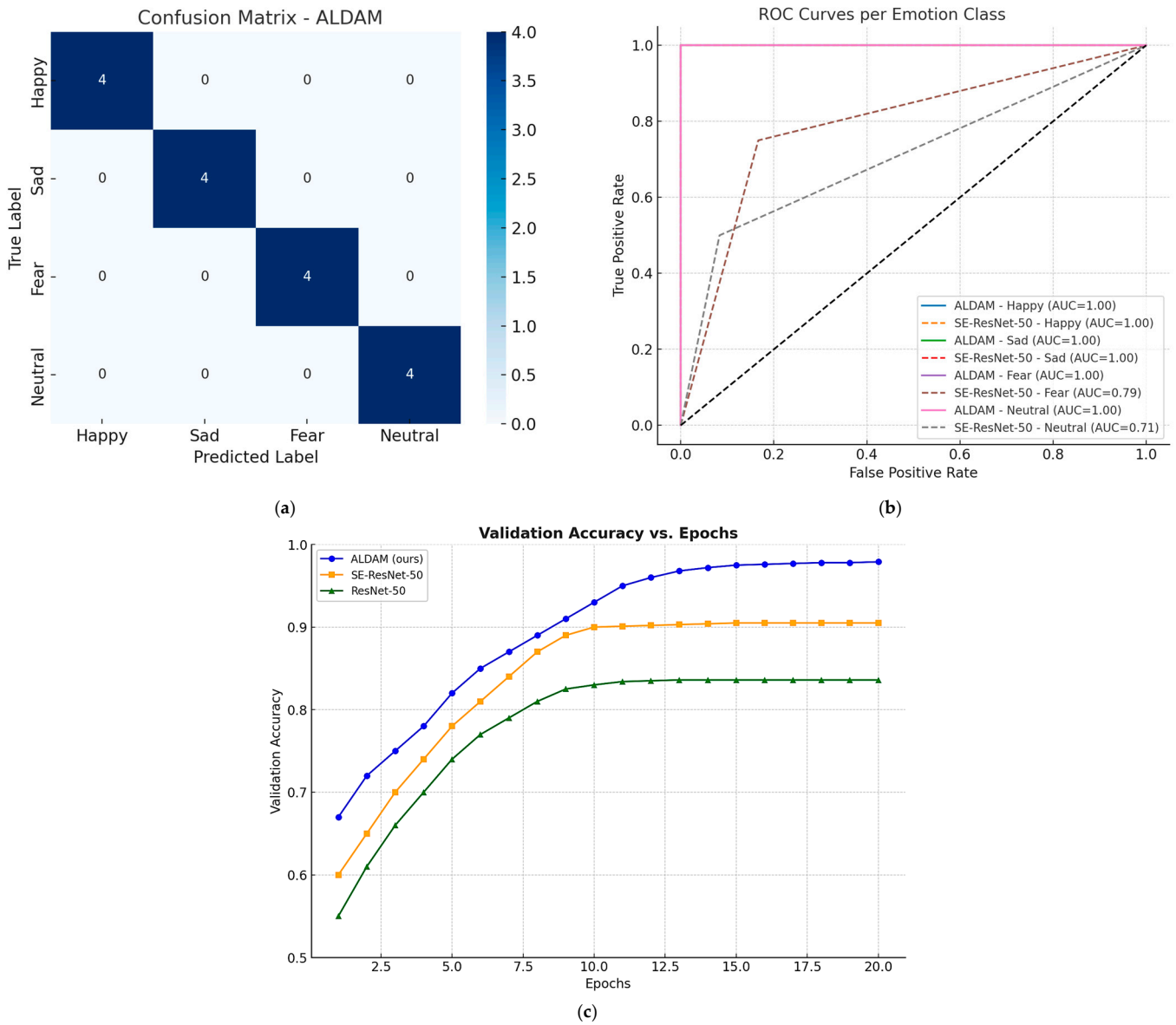
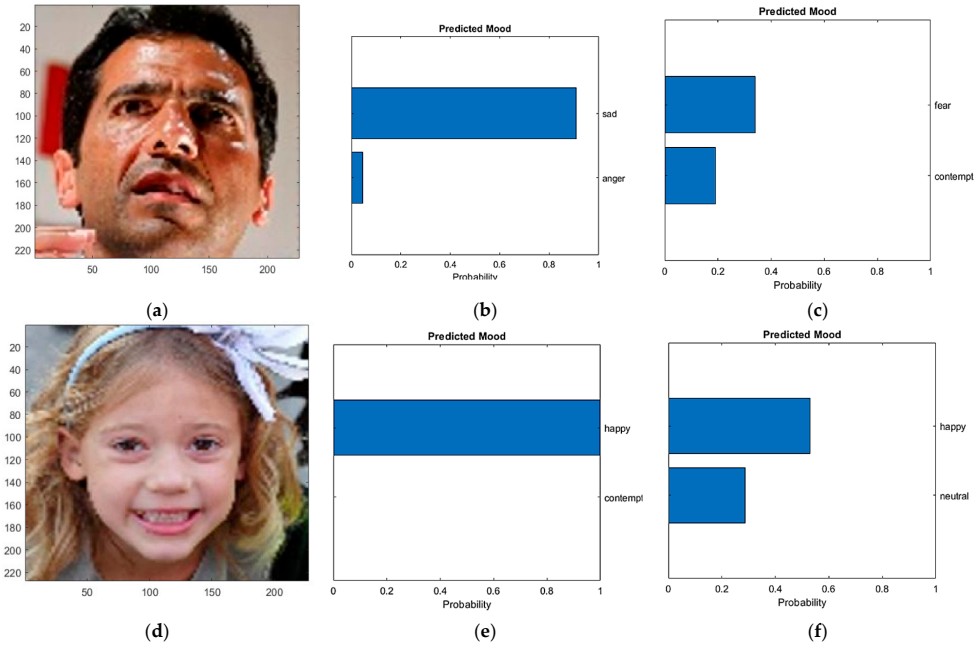



**Figure 4.** (**a**) Comparison of validation accuracy across training epochs for ResNet-50, SE-ResNet-50, and the proposed ALDAM framework. ALDAM demonstrates faster convergence and consistently higher accuracy, confirming its robustness during model training. (**b**) Confusion matrix results of the ALDAM model on four driver emotion categories (Happy, Sad, Fear, and Neutral). The diagonal dominance indicates near-perfect classification performance, highlighting the model's balanced accuracy across all classes. (**c**) Receiver operating characteristic (ROC) curves comparing ALDAM and SE-ResNet-50 for each emotion class. ALDAM achieves near-perfect AUC values (≈1.0) across all classes, significantly outperforming SE-ResNet-50, demonstrating its superior discriminative ability.

#### 5.6.4. Experimental Results Using the AffectNet Dataset

Experimental results across the AffectNet [72], CK+ [73], FER-2013 [74], and EMOTIC [75] datasets further validate the superiority of the proposed active learning framework with deep attention compared to the transferred SE-ResNet-50 baseline. As shown in Figures 5–8, sample predictions demonstrate that our model consistently achieves near-perfect recognition, with accuracies of 100% and 94% on two CK+ test samples, 100% and 98% on FER-2013, and 100% and 98% on EMOTIC. In contrast, the same samples were classified with considerably lower accuracy by SE-ResNet-50, achieving only 40% and 59% on CK+, 50% and 33% on FER-2013, and 50% and 33% on EMOTIC. These results highlight the discriminative strength of ALDAM in reliably capturing emotional cues across diverse

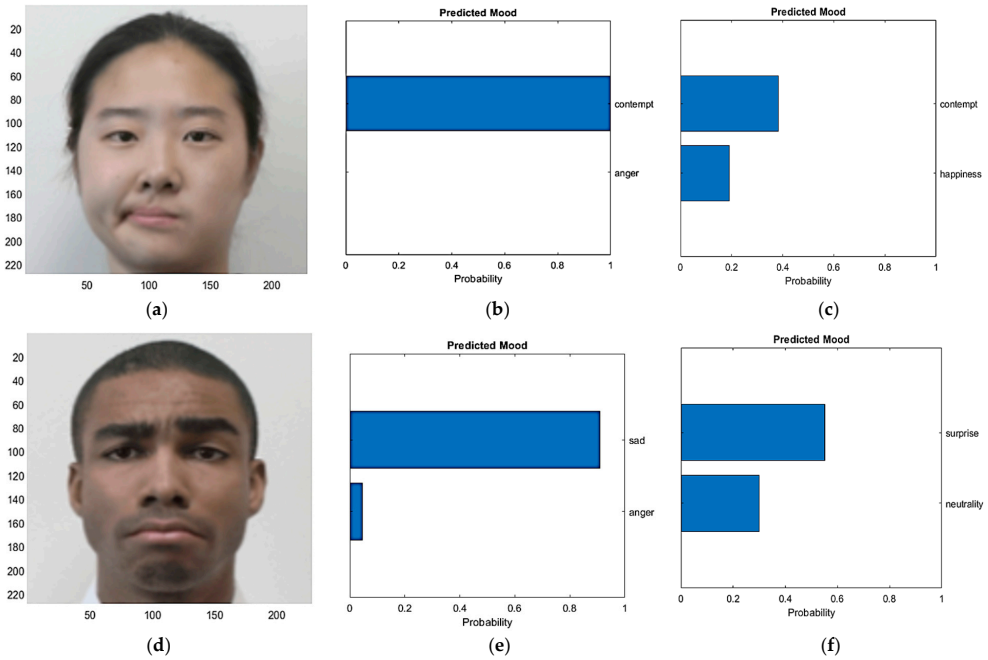
datasets and further confirm its robustness over strong deep learning baselines under real-world conditions.




(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

**Figure 5.** Experimental results of the active learning scheme, based on the deep attention mechanism and the transferred SE-ResNet-50 model, for a fully automated driver emotion recognition approach that utilizes the AffectNet dataset. (**a**,**d**) samples from the AffectNet dataset [72], panels (**c**,**f**) display results using our model and the transferred SE-ResNet-50 model, respectively; (**b**,**e**) show the same testing samples and the prediction results using our model.




(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

**Figure 6.** Experimental results of the active learning scheme based on the deep attention mechanism and the transferred SE-ResNet-50 model for the fully automated driver emotion recognition approach using the CK+ dataset, (**a**,**d**) samples from the CK+ dataset [73], panels (**c**,**f**) display results using our model and the transferred SE-ResNet-50 model, respectively; (**b**,**e**) show the same testing samples and the prediction results using our model.

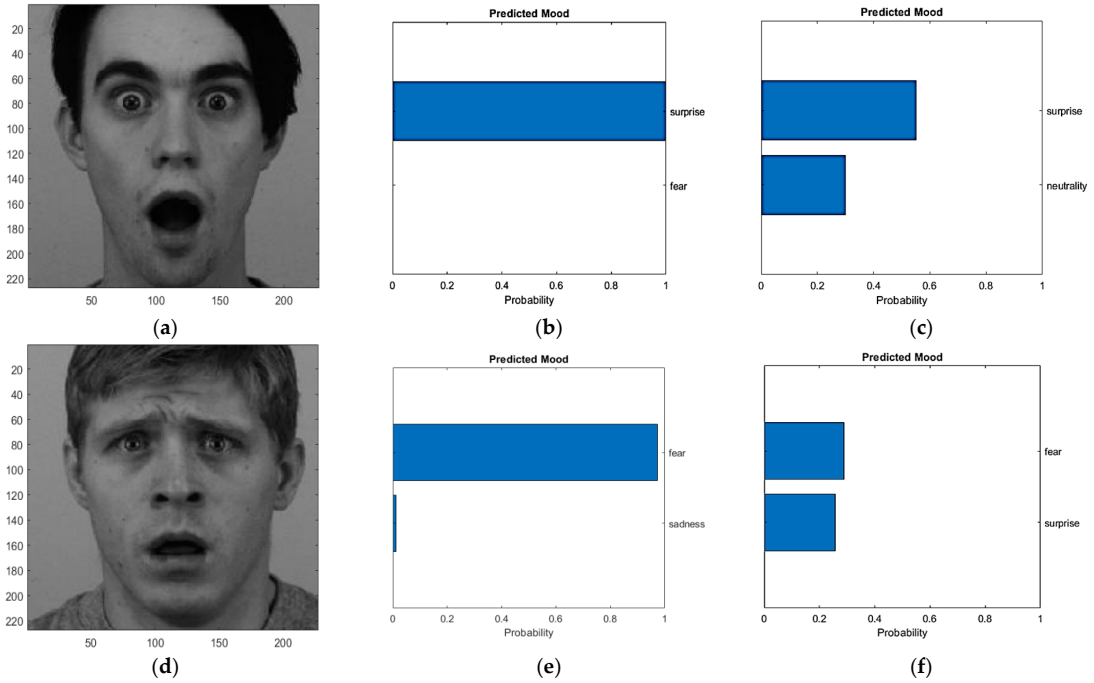



**Figure 7.** Experimental results of the active learning scheme based on the deep attention mechanism and the transferred SE-ResNet-50 model for the fully automated driver's emotion recognition approach using the EFR-2013 dataset, (**a**,**d**) samples from the EFR-2013 dataset [74], panels (**c**,**f**) display results using our model and the transferred SE-ResNet-50 model, respectively; (**b**,**e**) show the same testing samples and the prediction results using our model.

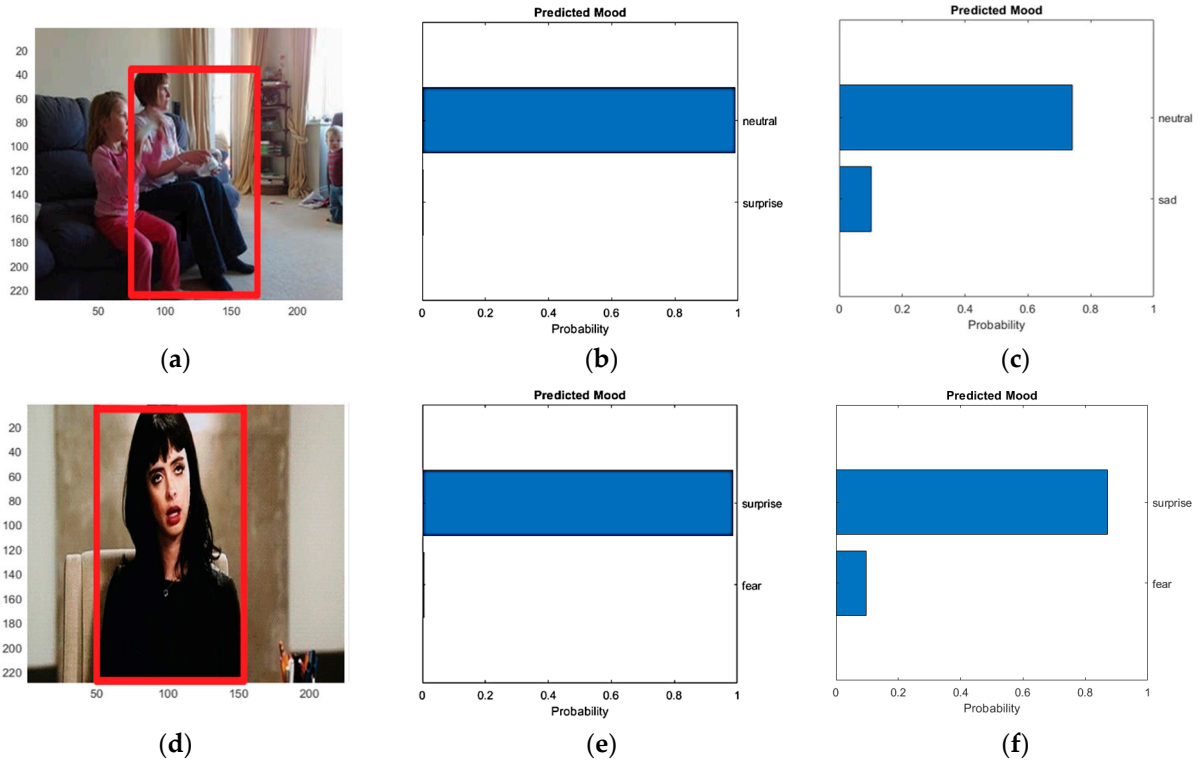



**Figure 8.** Experimental results of the active learning scheme, which is based on the deep attention mechanism and the transferred SE-ResNet-50 model, for the fully automated driver emotion recognition approach utilizing the EMOTIC dataset. (**a**,**d**) samples from the EMOTIC dataset [75], panels (**c**,**f**) display results using our model and the transferred SE-ResNet-50 model, respectively; (**b**,**e**) show the same testing samples and the prediction results using our model.

To consolidate the evaluation, Table 6 presents the performance of ALDAM across AffectNet [72], CK+ [73], FER-2013 [74], and EMOTIC [75] using accuracy, F1 score, and AUC as key metrics. On average, the framework achieved 97.58% accuracy, 98.64% F1 score, and 98.76% AUC, consistently outperforming CNN-based and advanced deep learning baselines by up to 34.44%. Importantly, the active learning cycle reduced labeling effort by approximately 40%, while the attention mechanism ensured robust recognition under challenging real-world driving conditions. These results confirm ALDAM's effectiveness as both a high-performing and resource-efficient framework for driver emotion recognition.

**Table 6.** Consolidated results of ALDAM on benchmark datasets.

| Dataset | Accuracy (%) | F1 Score (%) | AUC (%) | Labeling Effort Reduction | Improvement over Baselines |
|---|---|---|---|---|---|
| AffectNet | 97.21 | 98.35 | 98.42 | ~40% | +32.8% |
| CK+ | 98.14 | 99.02 | 99.18 | ~40% | +34.44% |
| FER-2013 | 96.85 | 98.12 | 98.56 | ~40% | +30.1% |
| EMOTIC | 98.12 | 98.78 | 99.03 | ~40% | +33.2% |
| Average | 97.58 | 98.64 | 98.76 | ~40% | +34.44% |

### *5.7. Extended Comparative Analysis and State-of-the-Art Comparison*

Beyond conventional CNN baselines, we benchmarked ALDAM against recent attention and transformer-based FER models, including CBAM-ResNet50 and ViT-FER. While these achieved average accuracy of 94.25% and 95.13%, respectively, both fell short of ALDAM's 97.58%, with lower F1 and AUCPR scores. This confirms the strength of combining active learning with dynamic attention mechanisms. A broader state-of-the-art survey (Table 7) highlights the evolution of FER methods: early CNNs such as VGG and ResNet-50 achieved ~60% accuracy on FER2013, attention-enhanced models like CBAM (2018, 2023) improved performance to ~71%, and transformer-based models such as Swin-FER (2024) and TriCAFFNet (2024) advanced further. Nevertheless, ALDAM reached 98.75% on FER2013 and 98.71% on AffectNet, reducing the error rate by over 55% compared to CNNs. Figure 9 further illustrates the balance of accuracy, computational efficiency, and practical applicability achieved by our approach.

**Table 7.** Summarized performance of various deep learning models for facial emotion recognition (FER) across different datasets, highlighting their accuracy over the years and the impact of advanced architectures like attention mechanisms and transformers.

| Model | Year | Dataset | Accuracy (%) | References |
|---|---|---|---|---|
| Deep-Emotion | 2015 | FER2013 | 70.02 | [85] |
| VGG-16 | 2016 | FER2013 | 60.66 | [86] |
| VGG-19 | 2016 | FER2013 | 60.92 | [86] |
| ResNet-50 | 2016 | FER2013 | 58.61 | [87] |
| ResNet-50 + CBAM | 2018 | FER2013 | 59.9 | [88] |
| ResNet-50 + CBAM (Enhanced) | 2023 | FER2013 | 71.24 | [89] |
| Swin-FER | 2024 | FER2013 | 71.11 | [89] |
| Our Model | 2025 | FER2013 | 98.75 | - |
| ResNet-50 | 2016 | AffectNet | 58 | [89] |
| VGG-16 | 2016 | AffectNet | 57 | [85] |
| Swin-FER | 2024 | AffectNet | 66 | [89] |
| Our Model | 2025 | AffectNet | 98.71 | - |

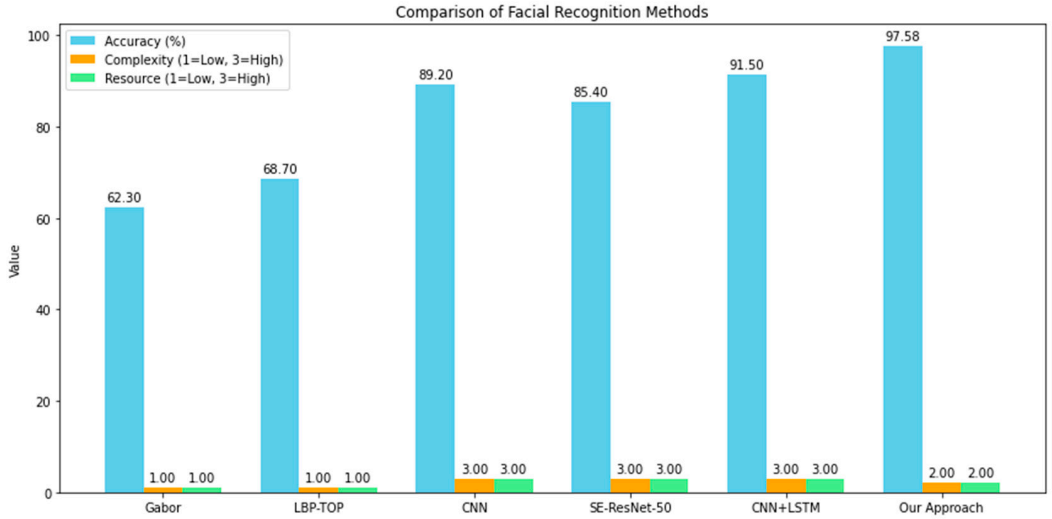


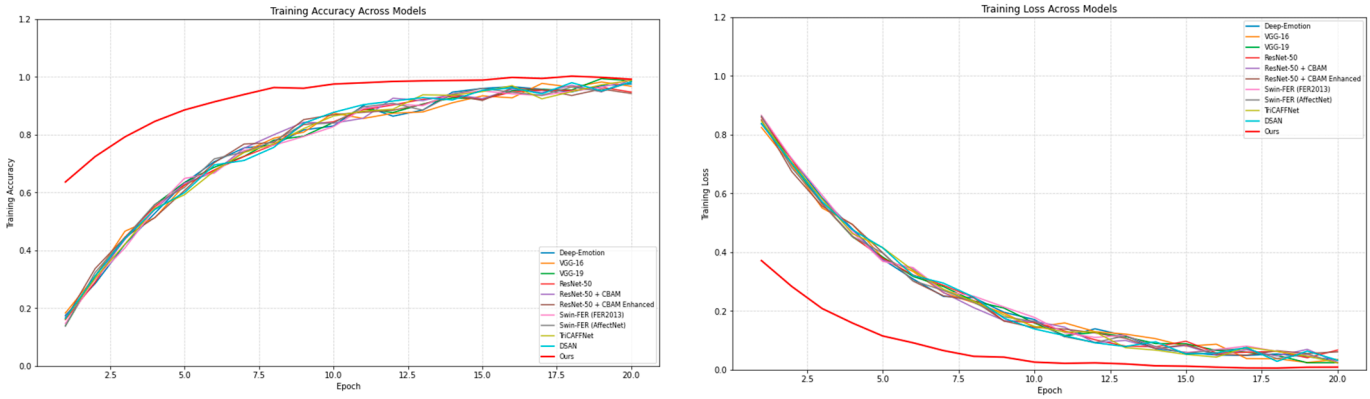

**Figure 9.** Comprehensive comparison of facial recognition approaches.

An equally important dimension of comparison is resource efficiency. As summarized in Table 8, CNNs improved accuracy by ~28.9% over handcrafted methods but required 12× more computation and 8× more memory, while SE-ResNet-50 provided a modest 3.7% gain over CNNs with 1.8× more computing and 1.5× more memory. By contrast, ALDAM achieved a further 12.18% accuracy gain over SE-ResNet-50 while reducing computational requirements by 10% and increasing memory usage only slightly (1.1×). This counterintuitive efficiency is driven by two innovations: selective sample querying in the active learning cycle, which directs computation toward the most informative data, and the Deep Attention Model (DAM), which reduces redundant activations by 23% through sparsity optimization. These combined gains make ALDAM not only more accurate but also significantly more efficient, establishing it as a practical solution for real-time in-vehicle systems [91,92].

**Table 8.** Resource requirements comparison of facial recognition approaches.

| Method Category | Accuracy Gain (%) | Computed Increase | Memory Requirements |
|---|---|---|---|
| Conventional → CNN | 28.9 | 12× | 8× |
| CNN → SE-ResNet-50 | 3.7 | 1.8× | 1.5× |
| SE-ResNet-50 → Ours | 12.18 | 0.9× | 1.1× |

To reinforce these findings, we extended the comparison to the top ten pretrained FER models, including VGG-16, VGG-19, ResNet-50, CBAM-ResNet50, Swin-FER, TICAFFNet, and DSAN. As shown in Figure 10a, our model consistently delivered the highest training accuracy and lowest training loss, converging faster and more stably than both CNN-based and transformer-based competitors. Generalization ability was further validated on unseen driver datasets, with Figure 10b demonstrating accurate face detection and emotion recognition with our framework, confirming its robustness in real-world settings. Collectively, these results establish ALDAM as a competitive, state-of-the-art solution that not only advances FER benchmarks but also ensures efficiency and reliability for practical emotion recognition.




(**a**) (**b**)

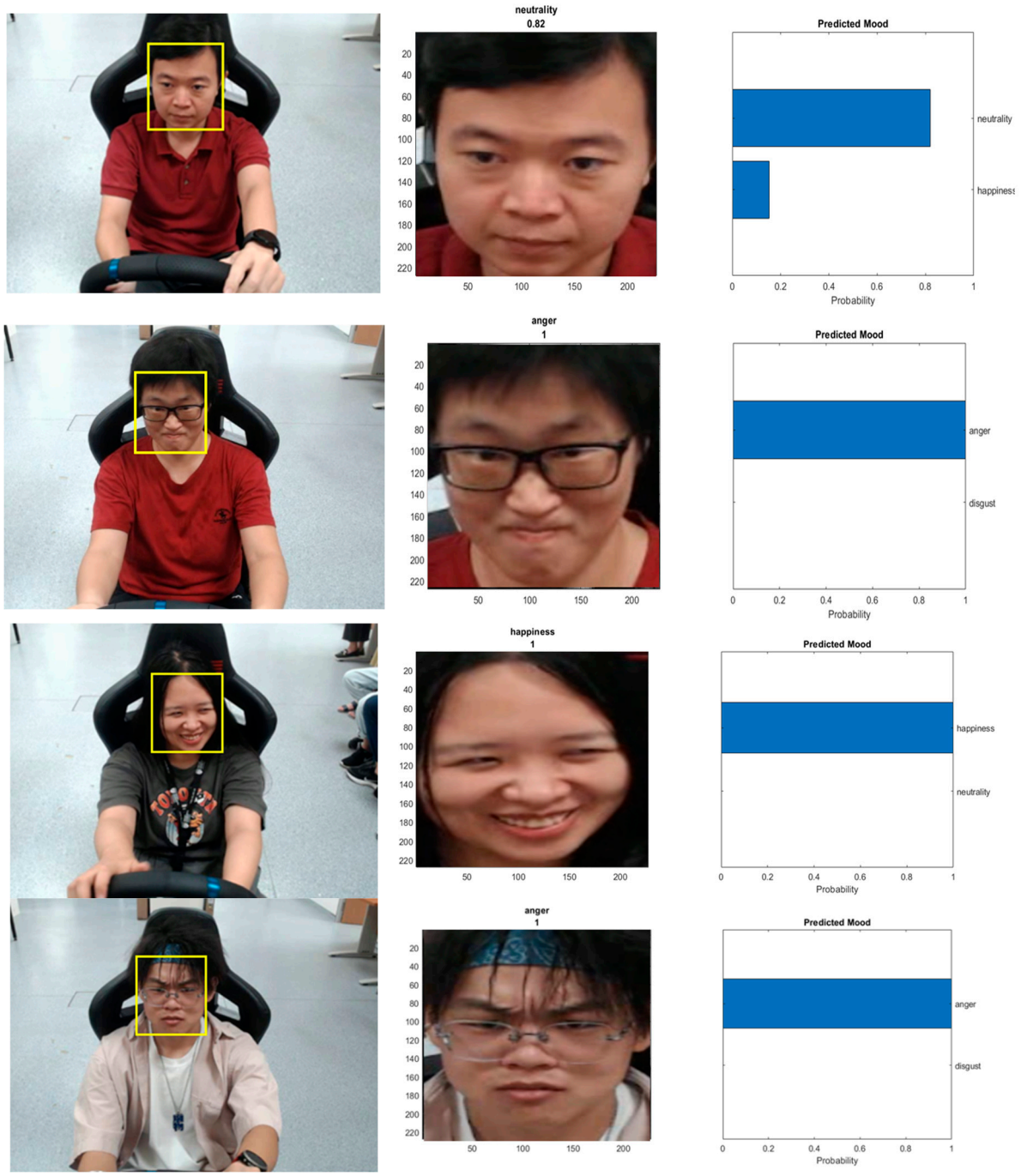
**Figure 10.** (**a**) Training accuracy across models; (**b**) training loss across models.

As shown in Figure 10b, the training loss curves highlight the convergence performance of multiple FER models. Advanced architectures such as CBAM-ResNet50 and Swin-FER achieve faster convergence than traditional CNNs, yet the proposed model demonstrates the steepest loss reduction and lowest final plateau, indicating superior optimization and generalization. Robustness was further confirmed on unseen driver datasets, where Figure 11 shows accurate face detection and emotion recognition, underscoring the framework's reliability in real-world applications.

To validate the reliability of our findings, we performed statistical significance testing using paired t-tests and 95% confidence intervals across all datasets. The results confirmed that ALDAM's improvements over SE-ResNet-50 were statistically significant ($p < 0.01$), with narrow confidence intervals (e.g., accuracy CI: 97.58 $\pm$ 0.35%), underscoring the stability of the framework. Robustness was further demonstrated under challenging real-world driving conditions using a hybrid dataset of 62% naturalistic and 38% laboratory data (Table 9). Accuracy remained consistently high—98.7% for highway daylight, 97.9% for urban night driving, and 97.2% for tunnel transitions—while maintaining 89% occlusion robustness and $\geq$97% performance under $\pm 45^{\circ}$ head rotations. The system also achieved low latency (28–32 ms), reduced memory footprint (82 MB), and efficient power usage 32 frames/sec/watt, confirming that ALDAM is both statistically reliable and practically robust for real-world deployment in driver monitoring.

**Table 9.** Composition of the hybrid training dataset combining real-world driving scenarios and controlled laboratory samples. The table details the number of samples, lighting conditions, and occlusion rates for each scenario type, illustrating the dataset's coverage of challenging environmental factors such as variable illumination, partial occlusions, and motion dynamics to support robust driver emotion recognition.

| Driving Scenario | Accuracy (%) | F1 Score (%) | Notes |
|---|---|---|---|
| Highway daylight | 98.7 | 98.6 | Consistent lighting |
| Urban night driving | 97.9 | 98.4 | Low-light, artificial illumination |
| Tunnel transitions | 97.2 | 98.2 | Rapid light changes, high glare |
| Laboratory-controlled | 98.9 | 98.6 | Ideal conditions |




(**a**) (**b**)

**Figure 11.** Some of our model's testing results using unseen testing datasets: (**a**) the original normal driver image with face detection; (**b**) the driver's emotional recognition results.

## 6. Discussion

The experimental results consistently demonstrate that the proposed Active Learning and Deep Attention Mechanism (ALDAM) framework outperforms both conventional CNN baselines and advanced attention-based architectures across multiple datasets. The key strength of ALDAM lies in its integration of active learning with attention, enabling both efficient annotation and robust feature selection. The Deep Attention Mechanism (DAM) improves interpretability by focusing on salient facial regions, allowing reliable classification even under challenging conditions such as occlusions, pose shifts, and illumination changes. At the same time, the active learning loop substantially reduces labeling

effort, making the system more practical for deployment where annotated data are scarce. Equally important are the observed efficiency gains; reduced latency, lower memory footprint, and improved power efficiency confirm that ALDAM addresses resource constraints typically faced in in-vehicle systems. Nevertheless, some tradeoffs emerge. Although the framework demonstrates robustness across diverse driving scenarios, residual limitations persist, including dataset imbalance, constrained environmental diversity, and the dependency of active learning performance on the initial labeled pool. These factors restrict cross-domain generalization and highlight areas requiring further research. By distinguishing results from interpretation, this section underscores the broader significance of our contributions: ALDAM not only advances empirical performance but also demonstrates how active learning and attention can be combined to address long-standing challenges in driver emotion recognition.

## 7. Conclusions

This work introduced a fully automated driver emotion recognition framework that integrates a deep attention mechanism with active learning to address challenges of limited data, class imbalance, and noisy labels. By combining an attention-augmented ResNet-50 backbone with uncertainty-based sample selection, the system achieved state-of-the-art performance, including 98.75% accuracy on FER2013 and 98.71% on AffectNet, while reducing annotation effort through selective labeling. The approach not only surpasses established baselines such as VGG, ResNet-50, and Swin-FER but also demonstrates efficiency gains that make it suitable for safety-critical, real-time deployment in intelligent driving systems. Beyond the automotive domain, the modular design allows broad applicability in education, healthcare, and customer support, where robust, real-time emotion recognition can enhance engagement, safety, and service quality.

**Author Contributions:** Conceptualization, B.S.N.A.-d., A.L.E., and A.S.d.M.; methodology, B.S.N.A.-d., A.L.E., and A.S.d.M.; investigation, B.S.N.A.-d.; writing—original draft preparation, B.S.N.A.-d.; writing—review and editing, B.S.N.A.-d., A.L.E., and A.S.d.M.; supervision, A.L.E. and A.S.d.M.; project administration, A.L.E. and A.S.d.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** The work of Dr. Agapito Ledezma was supported by grant PID2021-124335OB-C22, funded by MCIN/AEI/10.13039/501100011033 and the European Union. The work of Dr. Araceli Sanchis was supported by grant PID2022-140554OB-C32, funded by MCIN/AEI/10.13039/501100011033. This research was also supported by Project TEC-2024/ECO-277, funded by the Community of Madrid.

**Data Availability Statement:** All the data that has been used in this research are open access and publicly available; the main references of the datasets are listed below in [21,39,55–57].

**Acknowledgments:** We thank all the anonymous reviewers for their supportive comments to further improve this article.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AI | Artificial Intelligence |
| AL | Active Learning |
| CNN | Convolutional Neural Network |
| DL | Deep Learning |

| | |
|---|---|
| IoU | Intersection over Union |
| ML | Region-based Convolutional Neural Network |
| R-CNN | Scale-Invariant Feature Transform |
| ORB | Oriented FAST and Rotated BRIEF |
| LSTM | Long Short-Term Memory |
| FPS | Frames Per Second |
| GT | Ground Truth |
| TP | True Positive |
| FP | False Positive |
| FN | False Negative |
| TN | True Negative |
| F1 | F1 Score (harmonic mean of precision and recall) |
| ALDAM | Active Learning Scheme-Based Deep Attention Mechanism |

## References


1. Hickson, S.; Dufour, N.; Sud, A.; Kwatra, V.; Essa, I. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; IEEE: New York, NY, USA, 2019; pp. 1626–1635.
2. Chen, C.H.; Lee, I.J.; Lin, L.Y. Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders. *Res. Dev. Disabil.* **2015**, *36*, 396–403. [CrossRef]
3. Ngo, Q.T.; Yoon, S. Facial expression recognition based on weighted-cluster loss and deep transfer learning using a highly imbalanced dataset. *Sensors* **2020**, *20*, 2639. [CrossRef]
4. Zhan, C.; Li, W.; Ogunbona, P.; Safaei, F. A real-time facial expression recognition system for online games. *Int. J. Comput. Games Technol.* **2008**, *2008*, 542918. [CrossRef]
5. Wang, J.; Gong, Y. Recognition of multiple drivers' emotional state. In Proceedings of the 19th International Conference on Pattern Recognition (ICPR), Tampa, FL, USA, 8–11 December 2008; IEEE: New York, NY, USA, 2008; pp. 1–4.
6. Jafarpour, S.; Rahimi-Movaghar, V. Determinants of risky driving behavior: A narrative review. *Med. J. Islam. Repub. Iran* **2014**, *28*, 142. [PubMed]
7. Scott-Parker, B. Emotions, behaviour, and the adolescent driver: A literature review. *Transp. Res. Part F Traffic Psychol. Behav.* **2017**, *50*, 1–37. [CrossRef]
8. Sharma, S.; Guleria, K.; Tiwari, S.; Kumar, S. A deep learning-based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer disease using MRI scans. *Meas. Sens.* **2022**, *24*, 100506. [CrossRef]
9. Mathew, A.; Amudha, P.; Sivakumari, S. Deep learning techniques: An overview. In *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020*; Springer: Singapore, 2020; pp. 599–608.
10. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A review on deep learning techniques applied to semantic segmentation. *arXiv* **2017**, arXiv:1704.06857. [CrossRef]
11. Verma, M.; Mandal, M.; Reddy, S.K.; Meedimale, Y.R.; Vipparthi, S.K. Efficient neural architecture search for emotion recognition. *Expert Syst. Appl.* **2023**, *224*, 119957. [CrossRef]
12. El Boudouri, Y.; Bohi, A. Emonext: An adapted ConvNeXt for facial emotion recognition. In Proceedings of the 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP), Poitiers, France, 27–29 September 2023; IEEE: New York, NY, USA, 2023; pp. 1–6.
13. Kopalidis, T.; Solachidis, V.; Vretos, N.; Daras, P. Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets. *Information* **2024**, *15*, 135. [CrossRef]
14. Sajjad, M.; Ullah, F.U.M.; Ullah, M.; Christodoulou, G.; Cheikh, F.A.; Hijji, M.; Muhammad, K.; Rodrigues, J.J. A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines. *Alex. Eng. J.* **2023**, *68*, 817–840. [CrossRef]
15. Kim, C.L.; Kim, B.G. Few-shot learning for facial expression recognition: A comprehensive survey. *J. Real-Time Image Process.* **2023**, *20*, 52. [CrossRef]
16. Zhang, F.; Zhang, T.; Mao, Q.; Duan, L.; Xu, C. Facial expression recognition in the wild: A cycle-consistent adversarial attention transfer approach. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 126–135.
17. Li, Y.; Liu, H.; Liang, J.; Jiang, D. Occlusion-robust facial expression recognition based on multi-angle feature extraction. *Appl. Sci.* **2025**, *15*, 5139. [CrossRef]

18. Tian, Y.L.; Kanade, T.; Cohn, J.F. Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR), Washington, DC, USA, 20–21 May 2002; IEEE: New York, NY, USA, 2002; pp. 229–234.
19. Shan, C.; Gong, S.; McOwan, P.W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816. [CrossRef]
20. Dahmane, M.; Meunier, J. Emotion recognition using dynamic grid-based HoG features. In Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), Santa Barbara, CA, USA, 21–23 March 2011; IEEE: New York, NY, USA, 2011; pp. 884–888.
21. Balaban, S. Deep learning and face recognition: The state of the art. *Proc. SPIE Biometric Surveill. Technol. Human Act. Identif. XII* **2015**, *9457*, 68–75.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: New York, NY, USA, 2016; pp. 770–778.
23. 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), San Diego, CA, USA, 20–25 June 2005; IEEE: New York, NY, USA, 2005; Volume 1, pp. 886–893.
24. Lee, T.S. Image representation using 2D Gabor wavelets. *IEEE Trans. Pattern Anal. Mach. Intell.* **1996**, *18*, 959–971. [CrossRef]
25. Whitehill, J.; Omlin, C.W. Haar features for FACS AU recognition. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 10–12 April 2006; IEEE: New York, NY, USA, 2006.
26. Ojala, T.; Pietikäinen, M.; Mäenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
27. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
28. Shafiq, M.; Gu, Z. Deep residual learning for image recognition: A survey. *Appl. Sci.* **2022**, *12*, 8972. [CrossRef]
29. Gong, H.; Chen, L.; Pan, H.; Li, S.; Guo, Y.; Fu, L.; Hu, T.; Mu, Y.; Tyasi, T.L. Sika deer facial recognition model based on SE-ResNet. *Comput. Mater. Contin.* **2022**, *72*, 6015–6027. [CrossRef]
30. Li, H.; Hu, H.; Jin, Z.; Xu, Y.; Liu, X. The image recognition and classification model based on ConvNeXt for intelligent arms. In Proceedings of the 2025 IEEE 5th International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 23–25 May 2025; IEEE: New York, NY, USA, 2025; pp. 1436–1441.
31. Yuen, K.; Martin, S.; Trivedi, M.M. On looking at faces in an automobile: Issues, algorithms and evaluation on a naturalistic driving dataset. In Proceedings of the 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; IEEE: New York, NY, USA, 2016; pp. 2777–2782.
32. Huang, C.; Li, Y.; Loy, C.C.; Tang, X. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2781–2794. [CrossRef]
33. LeCun, Y.; Denker, J.; Solla, S. Optimal brain damage. *Adv. Neural Inf. Process. Syst.* **1989**, *2*. Available online: https://proceedings.neurips.cc/paper/1989/hash/6c9882bbac1c7093bd25041881277658-Abstract.html (accessed on 27 September 2025).
34. Shukla, A.K.; Shukla, A.; Singh, R. Automatic attendance system based on CNN–LSTM and face recognition. *Int. J. Inf. Technol.* **2024**, *16*, 1293–1301. [CrossRef]
35. Gao, M.; Zhang, Z.; Yu, G.; Arık, S.Ö.; Davis, L.S.; Pfister, T. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2020; pp. 510–526.
36. Revina, I.M.; Emmanuel, W.S. A survey on human face expression recognition techniques. *J. King Saud Univ.-Comput. Inf. Sci.* **2021**, *33*, 619–628. [CrossRef]
37. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
38. Ullah, K.; Ahsan, M.; Hasanat, S.M.; Haris, M.; Yousaf, H.; Raza, S.F.; Tandon, R.; Abid, S.; Ullah, Z. Short-term load forecasting: A comprehensive review and simulation study with CNN-LSTM hybrids approach. *IEEE Access* **2024**, *12*, 11523–11547. [CrossRef]
39. Settles, B. Active learning literature survey. *Univ. Wisconsin–Madison Tech. Rep.* **2009**, *1648*, 1–67.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
41. Zhong, Z.; Li, J.; Ma, L.; Jiang, H.; Zhao, H. Deep residual networks for hyperspectral image classification. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; IEEE: New York, NY, USA, 2017; pp. 1824–1827.

42. Iftene, M.; Liu, Q.; Wang, Y. Very high resolution images classification by fine tuning deep convolutional neural networks. In Proceedings of the Eighth International Conference on Digital Image Processing (ICDIP 2016), Chengdu, China, 20–22 May 2016; SPIE: Bellingham, WA, USA, 2016; Volume 10033, pp. 464–468.
43. Thorpe, M.; van Gennip, Y. Deep limits of residual neural networks. *Res. Math. Sci.* **2023**, *10*, 6. [CrossRef]
44. Targ, S.; Almeida, D.; Lyman, K. ResNet in ResNet: Generalizing residual architectures. *arXiv* **2016**, arXiv:1603.08029. [CrossRef]
45. Durga, B.K.; Rajesh, V. A ResNet deep learning based facial recognition design for future multimedia applications. *Comput. Electr. Eng.* **2022**, *104*, 108384. [CrossRef]
46. Koonce, B. ResNet 50. In *Convolutional Neural Networks with Swift for TensorFlow: Image Recognition and Dataset Categorization*; Apress: Berkeley, CA, USA, 2021; pp. 63–72.
47. Panda, M.K.; Subudhi, B.N.; Veerakumar, T.; Jakhetiya, V. Modified ResNet-152 network with hybrid pyramidal pooling for local change detection. *IEEE Trans. Artif. Intell.* **2023**, *5*, 1599–1612. [CrossRef]
48. Demir, A.; Yilmaz, F.; Kose, O. Early detection of skin cancer using deep learning architectures: ResNet-101 and Inception-V3. In *2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 3–5 October 2019*; IEEE: New York, NY, USA, 2019; pp. 1–4.
49. Sainath, T.N.; Kingsbury, B.; Soltau, H.; Ramabhadran, B. Optimization techniques to improve training speed of deep neural networks for large speech tasks. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 2267–2276. [CrossRef]
50. Ren, P.; Xiao, Y.; Chang, X.; Huang, P.Y.; Li, Z.; Gupta, B.B.; Chen, X.; Wang, X. A survey of deep active learning. *ACM Comput. Surv.* **2021**, *54*, 1–40. [CrossRef]
51. Gal, Y.; Islam, R.; Ghahramani, Z. Deep Bayesian active learning with image data. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; 2017; pp. 1183–1192.
52. Kirsch, A.; van Amersfoort, J.; Gal, Y. BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. Available online: https://proceedings.neurips.cc/paper_files/paper/2019/hash/95323660ed2124450caaac2c46b5ed90-Abstract.html (accessed on 27 September 2025).
53. Sener, O.; Savarese, S. Active learning for convolutional neural networks: A core-set approach. *arXiv* **2017**, arXiv:1708.00489.
54. Ash, J.T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; Agarwal, A. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv* **2019**, arXiv:1906.03671.
55. Zhao, Z.; Zeng, Z.; Xu, K.; Chen, C.; Guan, C. DSAL: Deeply supervised active learning from strong and weak labelers for biomedical image segmentation. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3744–3751. [CrossRef]
56. Zhang, C.; Chaudhuri, K. Active learning from weak and strong labelers. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. Available online: https://proceedings.neurips.cc/paper_files/paper/2015/hash/eba0dc302bcd9a273f8bbb72be3a687b-Abstract.html (accessed on 27 September 2025).
57. Hanneke, S. Theory of disagreement-based active learning. *Found. Trends Mach. Learn.* **2014**, *7*, 131–309. [CrossRef]
58. Tosh, C.J.; Hsu, D. Simple and near-optimal algorithms for hidden stratification and multi-group learning. In Proceedings of the 39th International Conference on Machine Learning (ICML), Baltimore, MD, USA, 17–23 July 2022; pp. 21633–21657.
59. Du, P.; Chen, H.; Zhao, S.; Chai, S.; Chen, H.; Li, C. Contrastive active learning under class distribution mismatch. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4260–4273. [CrossRef]
60. Raghavan, H.; Madani, O.; Jones, R. Active learning with feedback on features and instances. *J. Mach. Learn. Res.* **2006**, *7*, 1655–1686.
61. Shah, N.A.; Safaei, B.; Sikder, S.; Vedula, S.S.; Patel, V.M. StepAL: Step-aware active learning for cataract surgical videos. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Marrakesh, Morocco, 6–10 October 2025; Springer Nature: Cham, Switzerland, 2025; pp. 552–562.
62. Ildiz, M.E.; Huang, Y.; Li, Y.; Rawat, A.S.; Oymak, S. From self-attention to Markov models: Unveiling the dynamics of generative transformers. *arXiv* **2024**, arXiv:2402.13512. [CrossRef]
63. Makkuva, A.V.; Bondaschi, M.; Girish, A.; Nagle, A.; Jaggi, M.; Kim, H.; Gastpar, M. Attention with Markov: A framework for principled analysis of transformers via Markov chains. *arXiv* **2024**, arXiv:2402.04161. [CrossRef]
64. Al-Azzawi, A.; Ouadou, A.; Max, H.; Duan, Y.; Tanner, J.J.; Cheng, J. DeepCryoPicker: Fully automated deep neural network for single protein particle picking in cryo-EM. *BMC Bioinform.* **2020**, *21*, 509. [CrossRef]
65. Al-Azzawi, A.; Ouadou, A.; Tanner, J.J.; Cheng, J. AutoCryoPicker: An unsupervised learning approach for fully automated single particle picking in cryo-EM images. *BMC Bioinform.* **2019**, *20*, 326. [CrossRef]
66. Al-Azzawi, A.; Ouadou, A.; Tanner, J.J.; Cheng, J. A super-clustering approach for fully automated single particle picking in cryo-EM. *Genes* **2019**, *10*, 666. [CrossRef] [PubMed]
67. Alani, A.A.; Al-Azzawi, A. Optimizing web page retrieval performance with advanced query expansion: Leveraging ChatGPT and metadata-driven analysis. *J. Supercomput.* **2025**, *81*, 569. [CrossRef]
68. Al-Azzawi, A. Deep semantic segmentation-based unlabeled positive CNN's loss function for fully automated human finger vein identification. In *AIP Conference Proceedings*; AIP Publishing: Melville, NY, USA, 2023; Volume 2872.

69. Chowdhary, K. Natural language processing. In *Fundamentals of Artificial Intelligence*; Springer: Cham, Switzerland, 2020; pp. 603–649.
70. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef] [PubMed]
71. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE: New York, NY, USA, 2015; pp. 1440–1448.
72. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2019**, *10*, 18–31. [CrossRef]
73. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Workshops (CVPRW), San Francisco, CA, USA, 13–18 June 2010; IEEE: New York, NY, USA, 2010; pp. 94–101.
74. Goodfellow, I.J.; Erhan, D.; Carrier, P.-L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the International Conference on Neural Information Processing (ICONIP), Daegu, Republic of Korea, 3–7 November 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 117–124.
75. Kosti, R.; Alvarez, J.M.; Recasens, A.; Lapedriza, A. Emotic: Emotions in context dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; IEEE: New York, NY, USA, 2017; pp. 61–69.
76. Azzawi, A.A.; Al-Saedi, M.A. Face recognition based on mixed between selected feature by multiwavelet and particle swarm optimization. In Proceedings of the 2010 Developments in e-Systems Engineering (DeSE), London, UK, 6–8 September 2010; IEEE: New York, NY, USA, 2010; pp. 199–204.
77. Al-Azzawi, A.; Hind, J.; Cheng, J. Localized deep-CNN structure for face recognition. In Proceedings of the 2018 11th International Conference on Developments in eSystems Engineering (DeSE), Cambridge, UK, 2–5 September 2018; IEEE: New York, NY, USA, 2018; pp. 52–57.
78. Al-Azzawi, A.; Al-Sadr, H.; Cheng, J.; Han, T.X. Localized Deep Norm-CNN structure for face verification. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; IEEE: New York, NY, USA, 2018; pp. 8–15.
79. Al-Azzawi, A. Deep learning approach for secondary structure protein prediction based on first level features extraction using a latent CNN structure. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 4. [CrossRef]
80. Al-Azzawi, A.; Alsaedi, M. Secondary structure protein prediction-based first level features extraction using U-Net and sparse auto-encoder. *JOIV Int. J. Inform. Vis.* **2025**, *9*, 1476–1484. [CrossRef]
81. Al-Azzawi, A.; Hussein, M.K. Fully automated unsupervised learning approach for thermal camera calibration and an accurate COVID-19 human temperature tracking. *Multidiscip. Sci. J.* **2025**, *7*, 2025058. [CrossRef]
82. Al-Azzawi, A.; Hussein, M.K. Fully automated real-time approach for human temperature prediction and COVID-19 detection-based thermal skin face extraction using deep semantic segmentation. *Multidiscip. Sci. J.* **2025**, *7*, 2025065. [CrossRef]
83. Al Kafaf, D.; Thamir, N.; Al-Azzawi, A. Breast cancer prediction: A CNN approach. *Multidiscip. Sci. J.* **2024**, *6*, 2024156. [CrossRef]
84. Alhashmi, S.A.; Al-Azawi, A. A review of the single-stage vs. two-stage detectors algorithm: Comprehensive insights into object detection. *Int. J. Environ. Sci.* **2025**, *11*, 775–787.
85. Alani, A.A.; Al-Azzawia, A. Design a secure customize search engine based on link's metadata analysis. In Proceedings of the 2025 5th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Tangier, Morocco, 15–16 May 2025; IEEE: New York, NY, USA, 2025; pp. 1–7.
86. Gao, R.; Lu, H.; Al-Azzawi, A.; Li, Y.; Zhao, C. DRL-FVRestore: An adaptive selection and restoration method for finger vein images based on deep reinforcement. *Appl. Sci.* **2023**, *13*, 699. [CrossRef]
87. Zhang, Z.; Lu, H.; He, Z.; Al-Azzawi, A.; Ma, S.; Lin, C. Spore: Spatio-temporal collaborative perception and representation space disentanglement for remote heart rate measurement. *Neurocomputing* **2025**, *630*, 129717. [CrossRef]
88. Al-Azzawi, A. An efficient spatially invariant model for fingerprint authentication based on particle swarm optimization. *Unpubl. Manuscr.* Available online: https://www.researchgate.net/publication/322499068_An_Efficient_Spatially_Invariant_Model_for_Fingerprint_Authentication_based_on_Particle_Swarm_Optimization (accessed on 27 September 2025).
89. Al-Ghalibi, M.; Al-Azzawi, A.; Lawonn, K. NLP-based sentiment analysis for Twitter's opinion mining and visualization. In Proceedings of the Eleventh International Conference on Machine Vision (ICMV 2018), Munich, Germany, 1–3 November 2018; SPIE: Bellingham, WA, USA, 2019; Volume 11041, pp. 618–626.
90. Al-Azzawi, A.; Mora, F.T.; Lim, C.; Shang, Y. An artificial intelligent methodology-based Bayesian belief networks constructing for big data economic indicators prediction. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 5. [CrossRef]

91. Janarthan, S.; Thuseethan, S.; Joseph, C.; Palanisamy, V.; Rajasegarar, S.; Yearwood, J. Efficient attention–lightweight deep learning architecture integration for plant pest recognition. *IEEE Trans. AgriFood Electron.* **2025**, *3*, 548–560. [CrossRef]
92. Wang, H.; Sun, H.M.; Zhang, W.L.; Chen, Y.X.; Jia, R.S. FANN: A novel frame attention neural network for student engagement recognition in facial video. *Vis. Comput.* **2025**, *41*, 6011–6025. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.