A Comparative Review of Deep Learning Architectures for Emotion Recognition with Experimental Validation in Clinical Behavioral Monitoring

Armida P. Salazar * and Vladimir Y. Mariano

College of Computing and Information Technologies, National University, Manila, Philippines
Email: apsalazar@national-u.edu.ph (A.P.S.); vymariano@national-u.edu.ph (V.Y.M.)
*Corresponding author

***Abstract*—Deep learning has significantly advanced emotion recognition across visual, physiological, and multimodal data. Nevertheless, selecting appropriate architectures for real-world deployment remains challenging due to variations in dataset characteristics, computational constraints, and application requirements. This paper presents a comparative analysis of major deep learning architectures for emotion recognition, including Convolutional Neural Networks (CNNs), ResNet, MobileNet, Long Short-Term Memory (LSTM), Bidirectional LSTM, and Vision Transformers (ViTs), highlighting their strengths and limitations across heterogeneous data environments. Building upon the comparative review, an experimental study was conducted to examine the feasibility of a hybrid deep learning framework for detecting clinically observable affective states from temporally evolving facial video data in palliative care settings. The proposed approach integrates spatial and temporal representations and is evaluated using Leave-One-Video-Out validation to simulate real-world deployment on previously unseen individuals. Experimental results demonstrate stable learning behavior and meaningful behavioral prediction patterns despite moderate numerical accuracy caused by patient variability and dataset imbalance. Analysis indicates that hybrid architectures are effective in distinguishing stable versus expressive behavioral conditions, supporting the interpretation of affect recognition as behavioral condition awareness rather than categorical emotion classification. The study provides practical insights into architecture selection and demonstrates that hybrid deep learning systems can support continuous, non-intrusive behavioral monitoring as assistive analytical tools across healthcare and related application domains.**

***Keywords*—emotion recognition, hybrid deep learning, affective computing, temporal modeling, healthcare monitoring, vision transformers**

I. Introduction

Understanding human emotions through artificial intelligence has become a focal point of research in domains such as healthcare, education, and human-computer interaction. Emotion recognition systems utilise machine learning and deep learning techniques to interpret and respond to human emotional states, leveraging visual, auditory, textual, and physiological data. These advances enable more empathetic technologies that support mental health, enhance communication, and improve user experience.

Early approaches to automated emotion recognition relied on traditional machine learning and manually designed features. For instance, Support Vector Machines (SVMs) and decision trees were employed to classify emotions based on facial or physiological signals. These methods worked to a degree, but accuracy was limited. With the rise of deep learning, however, much stronger models have emerged—such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) with their Long Short-Term Memory (LSTM) variants, and more recently, Transformers. These models can automatically detect patterns in data rather than relying on hand-crafted features, leading to substantial improvements. For example, Yang *et al.* [1] reported that Electroencephalography (EEG)-based classifiers using CNNs and RNNs achieved about 81% accuracy for valence and 80% for arousal, outperforming SVMs (71.7%) and decision trees (69.4%). These advances highlight the ability of deep learning to surpass traditional approaches [1, 2].

The availability of large and diverse datasets has also driven progress. Facial Expression Recognition 2013 dataset (FER2013), introduced in a Kaggle competition, contains over 35,000 labeled grayscale facial images across seven emotion categories and has become a common baseline for training models [3]. AffectNet is one of the largest and most diverse datasets, comprising more than one million web-sourced images annotated for both categorical emotions and dimensional measures, such as valence and arousal, making it invaluable for building

Manuscript received January 28, 2026; revised March 11, 2026; accepted July 29, 2026; published September 14, 2026.

robust models [4]. The Cohn-Kanade (CK/CK+) corpus remains a gold standard for controlled experiments, offering high-quality recordings of both posed and spontaneous expressions [5]. Building on FER2013, FERPlus refines labels by using multiple annotators per image, thereby providing more reliable benchmarks for evaluating lightweight deep learning architectures, such as MobileNet [6, 7].

Beyond facial data, physiological signals also play an important role. The SJTU Emotion EEG (SEED) dataset, which records multichannel EEG signals, is widely used to benchmark CNN and LSTM models for temporal emotion recognition, offering strong baselines for studies in brain-signal analysis [8].

Our review focuses on six architectures—CNN, ResNet, MobileNet, LSTM, Bidirectional Long Short-Term Memory network (BiLSTM), and Vision Transformers—because they represent distinct milestones in the development of emotion recognition. CNNs provide the foundational baseline for visual affect analysis [1, 8], while ResNets extend this capability through very deep networks that capture subtle cues and achieve state-of-the-art results on benchmark datasets [9–11]. MobileNet introduces efficiency for real-time and resource-constrained environments [12–14]. LSTM and BiLSTM networks are included due to their strength in modeling temporal dynamics, which are critical for speech, text, and EEG signals [15–20]. Finally, Vision Transformers highlight the field's shift toward global self-attention mechanisms, demonstrating strong results on large-scale and complex datasets [21, 22]. More recently, transformer-based approaches have been evaluated using the Audio-Visual Facial Expression Recognition (AVFER) dataset, which was specifically designed to capture diverse visual emotion cues for Vision Transformer architectures [23].

Other architectures, such as Generative Adversarial Network (GAN) or Capsule Networks, remain less established in benchmark evaluations, so concentrating on these six models provides a balanced spectrum of accuracy, efficiency, and generalizability across modalities.

Despite these advances, systematically comparing these models remains necessary to clarify their relative strengths, limitations, and most effective application areas. In the following sections, we review the major deep learning architectures for emotion recognition, including CNNs (such as ResNet and MobileNet), LSTM and BiLSTM recurrent networks, and Vision Transformers. Their performance is evaluated across common metrics (accuracy, precision, recall, F1-score) and modalities (vision, speech, EEG, text).

The development of deep learning architectures for emotion recognition reflects the broader evolution of artificial intelligence from handcrafted feature engineering toward end-to-end representation learning. Early emotion recognition systems relied heavily on convolution-based architectures, particularly CNNs, which became dominant between 2015 and 2018 because of their ability to automatically extract hierarchical facial features from image data. Compared with traditional machine learning approaches, CNNs significantly improved recognition accuracy and established a new foundation for visual affective computing [9–11].

Between 2018 and 2020, research expanded beyond static image analysis to include temporal emotion modeling through Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) networks. These architectures enabled effective analysis of speech, physiological signals, textual content, and video sequences where emotional information evolves over time rather than appearing as isolated events [15–20].

The introduction of Vision Transformers (ViTs) from 2020 onwards marked a major architectural transition. By leveraging self-attention mechanisms, ViTs are capable of capturing global contextual relationships across facial regions and have demonstrated strong performance on large-scale and diverse emotion recognition datasets [21, 23].

II. Deep Learning Architectures for Emotion Recognition

A. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) serve as the fundamental architecture for visual emotion recognition due to their ability to learn hierarchical spatial features from facial images. Their layered convolutional structure enables the extraction of low-level information, such as edges and textures, followed by more abstract patterns, including facial muscle configurations associated with specific emotions [24–29]. CNNs have shown consistently strong performance on structured datasets, especially those captured under controlled conditions, where illumination, pose, and background variations are minimal. These stability conditions allow CNN filters to learn distinctive patterns that reliably separate emotional classes.

Beyond facial imagery, CNN-based approaches have been successfully adapted for other modalities, such as applying convolutions to spectrograms in speech emotion recognition [30, 31]. However, the standard CNN architecture is limited by its localized receptive fields, which may constrain its ability to capture global spatial relationships across disparate facial regions. In scenarios involving occlusion, spontaneous expressions, or in-the-wild conditions with substantial variability, CNNs may require additional regularization, deeper layers, or hybrid enhancements to maintain performance. These limitations motivated the development of deeper CNN variants and complementary architectures.

B. Residual Networks (ResNet)

Residual Networks (ResNets) enhance the capabilities of traditional CNNs by incorporating skip connections that enable information to bypass certain layers during training [9–11]. This mechanism mitigates vanishing gradients, allowing for the successful training of significantly deeper networks. In emotion recognition, deeper architectures translate to improved extraction of fine-grained spatial cues, allowing the model to detect subtle differences in facial regions that correspond to nuanced emotional states. As demonstrated in prior work,

ResNet-50 achieves high accuracy on structured datasets such as CK+ [32], indicating its capacity to generalize well on clean data with minimal noise.

Despite these strengths, ResNets require substantial computational resources and careful training to avoid overfitting, especially when dealing with small or imbalanced datasets. Their performance can be highly sensitive to dataset diversity, with notable degradation when applied to in-the-wild datasets characterized by variation in lighting, demographics, facial orientation, and image quality. These challenges highlight the architectural trend toward models that strike a balance between representational power and efficiency, as well as the adoption of global attention mechanisms seen in more recent architectures.

C. MobileNet

MobileNet addresses the demand for efficient, low-latency emotion recognition by introducing depthwise separable convolutions, which greatly reduce computational complexity and parameter count [12–14]. This design makes MobileNet particularly suitable for deployment in real-time systems such as mobile devices, IoT-based emotion sensing platforms, and classroom monitoring tools where energy efficiency and response time are essential [14].

Studies show that MobileNet performs competitively even on challenging datasets, such as FERPlus, achieving an accuracy of 88.11% [6], which demonstrates strong generalization under naturalistic conditions.

However, MobileNet's lightweight nature imposes limits on representational capacity, reducing its ability to model highly subtle emotional cues or complexities present in datasets with significant inter-class variability. While it excels in latency-sensitive applications, its performance typically falls short of deeper architectures on large-scale datasets with diverse expressions. MobileNet is therefore ideal for embedded deployment and rapid inference scenarios, but less optimal for high-stakes analytical or research settings requiring extensive emotional nuance modeling.

D. Long Short-Term Memory Networks (LSTM)

Long Short-Term Memory (LSTM) networks are specifically designed for processing sequential data and excel in emotion recognition from modalities such as speech, EEG, and temporal sensor readings. Their internal gating mechanisms enable LSTMs to capture and retain long-term dependencies, allowing for the modeling of emotional trajectories rather than instantaneous states [15–17]. This makes LSTMs particularly well-suited for domains where emotions unfold over time, such as speech emotion recognition or physiological signals that reflect dynamic changes in affective states.

Nevertheless, LSTMs are not inherently capable of modeling spatial relationships and often require integration with CNNs for image- or video-based emotion analysis. They also incur higher computational costs compared with simpler recurrent models, especially when processing long sequences. Their reliance on consistent temporal patterns can also make them sensitive to noise or irregular sampling. Despite these limitations, LSTMs are indispensable for modeling sequential emotions and play a central role in many multimodal frameworks [33–35].

E. Bidirectional LSTM (BiLSTM)

Bidirectional LSTM (BiLSTM) networks extend the capabilities of standard LSTMs by processing input sequences in both forward and backward directions, enabling the model to incorporate full contextual information from an entire sequence [18–20]. This bidirectional structure is particularly valuable for text-based emotion recognition, where understanding sentiment often requires interpreting both preceding and subsequent phrases. BiLSTMs have also proven effective in multimodal emotion recognition, especially when integrated with CNNs or attention mechanisms to combine spatial and temporal information [34–36].

However, because BiLSTMs require complete sequences, they are most effective in offline analysis rather than real-time applications where data arrives incrementally. They also have higher computational demands than unidirectional LSTMs, resulting in increased processing time and memory usage. Despite these constraints, BiLSTMs remain a powerful choice for tasks where contextual understanding is crucial for accurate emotion interpretation.

F. Vision Transformers (ViTs)

Vision Transformers (ViTs) represent a significant shift from convolution-based architectures by applying self-attention mechanisms to image patches, enabling the modeling of global dependencies across the entire image. This global attention enables ViTs to capture subtle interactions among distant facial regions, resulting in improved recognition of complex or ambiguous emotions. ViTs have demonstrated strong performance on large-scale datasets such as AffectNet and AVFER, benefiting from higher dataset diversity and sample richness [21].

However, the performance of ViTs heavily depends on the availability of large annotated datasets. Their patch-based self-attention mechanism lacks the strong inductive biases inherent in CNNs, which means ViTs may struggle with small datasets or highly noisy environments without additional regularization or hybridization. They also require substantial computational resources, limiting their applicability for real-time or edge-based systems. Nonetheless, ViTs represent a powerful family of models for advanced emotion recognition research.

III. Comparative Analysis and Discussion

A. Interpretation of Performance Across Datasets

The values presented in Table I reveal clear distinctions in how various deep learning architectures perform across heterogeneous emotion datasets, each with distinct modality characteristics, acquisition conditions, and complexity levels. For structured visual datasets such as CK+, ResNet-50 achieves a high accuracy of 97.60%, while its hybrid variant ResNet-50 + Gated Recurrent Unit (GRU) also performs strongly at 95.56%. These high values reinforce that deep convolutional architectures excel

when facial expressions are posed, lighting conditions are controlled, and noise is minimal. The stability of CK+ enables deeper models to exploit fine-grained facial cues—a trend consistent with reported results in earlier studies [32, 37].

For EEG-based emotion recognition, the CNN model reaches 89.50% accuracy with an F1-score of 0.85, outperforming LSTM, which achieves 86.48%. This performance gap highlights the advantage of CNN-based spectral feature extraction over purely temporal modeling when processing EEG signals from the SEED dataset. The relatively high accuracy for both models indicates that well-preprocessed EEG features provide strong discriminative power for emotion classification, consistent with prior research using SEED [8].

A different trend emerges for in-the-wild or complex visual datasets. For instance, the Vision Transformer (ViTFER) achieves an accuracy of only 50.05–53.10% and an F1-score of 0.6220 on the AVFER dataset. The substantial drop in accuracy relative to CNN and ResNet models on CK+ highlights the challenges posed by spontaneous expressions, varied lighting conditions, and cultural differences. Transformers require large, diverse datasets to learn robust attention patterns; AVFER's complexity and the naturalistic variability pose challenges for ViT models.

MobileNet demonstrates performance that reflects its balance between efficiency and accuracy. On FERPlus, MobileNet-V2 achieves an accuracy of 88.11%, while variants trained on FER+ and CAFE reach 89.00% and 65.11% accuracy, respectively. The spread in these values highlights how MobileNet's lightweight design adapts differently depending on dataset difficulty: it performs strongly on FERPlus, which has substantial sample diversity, but struggles on CAFE where expressions are subtler and lighting conditions vary. The CAFE model's F1-score of 0.6419 and latency of 45 ms further confirm that MobileNet is well-suited for real-time applications, even when accuracy declines under challenging conditions.

Overall, Table I shows that no single architecture dominates across all data types. CNNs and ResNets excel in structured facial datasets, MobileNet performs efficiently under operational constraints, ViTs struggle with small or highly variable datasets, and LSTM variants provide competitive performance in temporal modalities. These differences justify the need for architecture-specific selection logic based on dataset characteristics and application requirements.

TABLE I. DEEP LEARNING ARCHITECTURES PERFORMANCE ACROSS HETEROGENEOUS EMOTION DATASETS

Model & Dataset	Accuracy (%)	F1-score	Latency (ms)	Source
CNN (EEG, SEED Dataset)	89.50	0.85	35	[8]
LSTM (EEG, SEED Dataset)	86.48	—	—	[8]
Vision Transformer—ViTFER (AVFER Dataset)	50.05–53.10	0.6220	—	[21]
ResNet-50 (CK+ Dataset)	97.60	—	—	[32]
ResNet-50 + GRU (CK+ Dataset)	95.56	—	—	
MobileNet-V2 (FERPlus Dataset)	88.11	—	—	[37]
MobileNet-V2 (CAFE Dataset)	65.11	0.6419	45	
MobileNet (FER+ variant)	89.00	0.87	—	[6]

B. Dataset Difficulty Stratification

Stratifying datasets by difficulty provides deeper insight into the interplay between dataset characteristics and model performance. CK+ is categorized as a low-difficulty dataset due to its controlled environment, consistent lighting, and posed expressions, enabling deep CNN-based models to achieve near-perfect accuracy [32]. FERPlus falls into the medium-difficulty category, as it includes spontaneous expressions, varied illumination, and greater demographic diversity, challenging feature extraction and generalization [6].

High-difficulty datasets such as AffectNet and AVFER [21] present spontaneous, culturally diverse, and context-rich emotional displays that demand robust global modeling capabilities. ViTs often perform better in such settings when trained on large datasets. For modality-specific datasets, such as SEED EEG [8], the challenges lie in the inherent physiological variability of the signals. This difficulty classification system helps clarify why specific models excel under particular dataset conditions and supports more informed model selection.

In addition to modality differences, dataset complexity can be categorized according to environmental conditions and emotional variability. Controlled datasets such as CK+ contain standardized lighting conditions, frontal facial views, and posed emotional expressions, resulting in relatively low intra-class variation and high recognition accuracy. In contrast, in-the-wild datasets such as FERPlus and AffectNet contain spontaneous expressions, varying illumination conditions, pose variations, and partial occlusions that significantly increase classification difficulty [4, 6].

Cross-cultural datasets introduce additional challenges because emotional expression patterns may vary across demographic and cultural groups. Consequently, models trained primarily on one population may experience reduced generalization when evaluated on more diverse datasets. These observations explain why high-performing models on controlled datasets may exhibit substantially lower performance under realistic deployment conditions.

C. Model-Scenario-Data Matrix

The model-scenario-data matrix presented in Table II highlights the practical alignment between architectures and their ideal deployment settings. CNNs and ResNets are well-suited for environments with high-quality image inputs, such as laboratory-based studies or clinical screening applications. Their deep feature hierarchies enable robust classification when facial features are clearly visible and consistently captured. Conversely, MobileNet's

efficiency makes it ideal for real-time applications in mobile or IoT devices, where computational constraints are a limiting factor [12–14].

For modalities involving sequential data—such as speech, EEG, or multimodal signals—LSTMs and BiLSTMs offer substantial advantages due to their ability to model temporal dependencies [15–18, 34–36]. Vision Transformers, although computationally demanding, are best suited for large-scale, diverse datasets that require global contextual reasoning across facial regions. Hybrid models combine the strengths of these architectures, enabling more comprehensive emotion recognition in complex environments. This matrix underscores the importance of selecting models based on scenario-specific requirements rather than relying solely on accuracy benchmarks.

D. Comparative Analysis of Common Emotion Recognition Datasets

Dataset selection significantly influences model performance, generalization capability, and deployment suitability. Differences in data scale, modality, annotation strategy, demographic diversity, and environmental complexity contribute directly to variations in recognition accuracy across studies.

Table III summarizes the characteristics of the most widely used emotion recognition datasets, highlighting their approximate size, modality, diversity, annotation type, principal limitations, and recommended application scenarios. This comparison provides practical guidance for selecting datasets that align with specific research objectives and deployment requirements.

Dataset selection should align with application requirements. CK+ remains valuable for controlled benchmarking but is less suitable for evaluating cross-cultural generalization due to its limited diversity and posed emotional expressions [32]. FERPlus and AffectNet provide more realistic evaluation environments because they contain substantial variability in demographics, facial appearance, lighting conditions, and expression intensity [4, 6]. For physiological emotion recognition, SEED remains one of the most widely adopted benchmark datasets [8]. AVFER, meanwhile, supports the evaluation of multimodal and transformer-based architectures under more complex affective conditions [21].

TABLE II. DEEP LEARNING ARCHITECTURES AND DEPLOYMENT

Model Family	Most Suitable Application Scenarios	Dataset / Data Characteristics Where the Model Performs Best
CNN / ResNet	High-accuracy visual emotion recognition; controlled or semi-controlled environments; clinical or laboratory settings	Clear frontal facial images; structured datasets with minimal noise (e.g., CK+); medium-to-large image datasets where spatial hierarchies are prominent
MobileNet	Real-time mobile and IoT deployments; classrooms, telehealth monitoring, embedded systems requiring low latency	In-the-wild facial images with variability; small-to-medium datasets; resource-constrained environments where computational efficiency is required
LSTM	Speech-based emotion recognition; EEG/physiological emotion modeling; temporal affect analysis	Long sequential datasets such as speech waveforms or EEG recordings; data where emotional patterns evolve over time
BiLSTM	Text-based emotion recognition; multimodal fusion (audio + text; EEG + facial cues); offline sequence modeling	Complete sequential datasets that require contextual understanding from both past and future signals
Vision Transformer (ViT)	Large-scale visual datasets for research; spontaneous in-the-wild emotion recognition; datasets with complex spatial dependencies	High-diversity, large annotated datasets (e.g., AffectNet, AVFER); images requiring global attention modeling and cross-region relational learning
Hybrid Models (CNN-LSTM, CNN-ViT)	Video-based emotion recognition; multimodal affective computing; scenarios where spatial and temporal cues must be jointly modeled	Heterogeneous datasets combining frames, sequences, or multiple modalities; tasks requiring simultaneous spatial and temporal analysis

TABLE III. COMPARATIVE CHARACTERISTICS OF COMMON EMOTION RECOGNITION DATASETS

Dataset	Approximate Size	Modality	Diversity	Annotation Type	Main Limitation	Recommended Application
CK+	~600 sequences	Facial Video	Low	Basic emotion categories	Posed expressions	Controlled benchmarking
FER2013	35,000+ images	Facial Images	Medium	Seven emotion classes	Label noise	General FER training
FERPlus	35,000+ images	Facial Images	Medium	Crowd-annotated labels	Limited demographic diversity	Real-world FER benchmarking
AffectNet	>1 million images	Facial Images	High	Emotion and valence-arousal	Annotation variability	Large-scale training
SEED	EEG signals	Physiological	Medium	Emotional states	Limited subject population	EEG emotion recognition
AVFER	Audio-Visual	High	High	Multimodal affective labels	High computational demand	Transformer-based evaluation

E. Model Selection Decision Logic

The decision logic illustrated in Fig. 1 translates comparative insights into a practical tool for selecting appropriate architectures. Modality serves as the first decision point, since image-based tasks demand spatial modeling, whereas sequential modalities require architectures that capture temporal dependencies. Deployment constraints form the second decision point; MobileNet, for example, is a natural choice for resource-limited systems due to its efficient design [12–14], while deeper networks or ViTs may be more suitable for desktop or cloud-based deployments.

Data size and diversity constitute the third major decision factor. Large-scale datasets with high variability benefit from the global modeling capabilities of ViTs [21, 22], whereas smaller datasets with more uniform samples are better served by CNN or ResNet architectures [32]. Temporal dependencies provide the final

selection criterion, guiding users toward LSTM-based networks when emotions unfold over time. Together, these decision points form a systematic framework for choosing architectures tailored to real-world constraints.

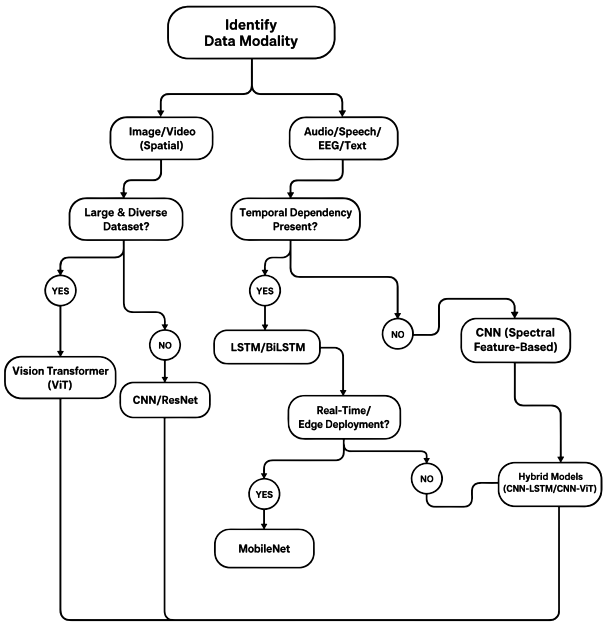

Fig. 1. Architecture selection.

For example, MobileNet is particularly suitable for real-time mobile emotion recognition applications due to its low computational requirements. Hybrid ResNet–Vision Transformer architectures may be preferable for healthcare applications involving subtle facial behavior analysis and micro-expression detection. For multimodal sentiment analysis involving text, speech, and facial cues, CNN-BiLSTM architectures offer a balanced approach for integrating spatial and temporal information.

IV. Performance and Model Complexity

Deep learning models for emotion recognition vary significantly in complexity, influencing their capability, resource requirements, and deployment feasibility. CNNs offer a balanced combination of efficiency and representational power, making them an excellent baseline for image-based tasks. ResNets build upon this by enabling deeper architectures that capture more intricate features, improving performance on datasets with subtle emotional cues. Meanwhile, LSTM and BiLSTM architectures extend deep learning's capability to sequential domains, enabling the modeling of dynamic emotional transitions in speech and physiological data.

Vision Transformers represent a shift toward global feature modeling, leveraging self-attention to capture long-range interactions across facial regions. However, this increase in complexity comes with higher resource requirements and a need for large datasets to avoid overfitting. Hybrid models, which combine the strengths of CNNs, LSTMs, and Transformers, illustrate the field's progression toward architectures capable of handling multimodal and multi-context emotion data. These trends highlight the importance of considering model complexity in relation to both application requirements and dataset characteristics.

A. Practical Model Optimization Techniques

Beyond architecture selection, practical deployment often requires optimization strategies that improve efficiency, scalability, and robustness. For lightweight architectures such as MobileNet, model pruning, quantization, and compression techniques can significantly reduce memory requirements and inference latency while preserving competitive recognition performance. These optimizations are particularly beneficial for mobile devices, edge computing platforms, and IoT-based emotion recognition systems [12–14].

Vision Transformers commonly benefit from transfer learning strategies that leverage large-scale pretraining before domain-specific fine-tuning. Such approaches help mitigate the substantial data requirements associated with self-attention mechanisms and improve performance when labeled emotion datasets are limited [21, 23].

For LSTM and BiLSTM architectures, attention mechanisms have been widely adopted to emphasize emotionally relevant temporal segments and improve long-sequence modeling. These approaches reduce the influence of irrelevant temporal information and enhance recognition performance across sequential modalities [18–20]. Hybrid architectures may further improve performance through feature-level fusion strategies that integrate spatial and temporal representations within a unified learning framework.

B. Open-Source Development Ecosystem

The growing availability of open-source frameworks and deployment tools has accelerated research and practical implementation in emotion recognition systems. Commonly used tools include OpenCV for face detection, PyTorch Lightning for model development, ONNX for cross-platform deployment, and FER libraries for rapid prototyping. Recent studies further demonstrate the effectiveness of these ecosystems in supporting advanced model implementation and deployment. For example, Shanimol and Charles [37] implemented a hybrid ResNet50–GRU model for facial emotion recognition, while Banerjee *et al.* [38] demonstrated efficient training and profiling of emotion recognition models on mobile devices. Collectively, these open-source tools and implementation frameworks have significantly advanced reproducible research and real-world deployment in affective computing.

V. Hybrid and Ensemble Approaches

Hybrid architectures integrate the strengths of complementary models, addressing limitations inherent in single-architecture approaches. CNN-LSTM hybrids, for example, combine spatial feature extraction with temporal modeling, making them particularly effective for video-based emotion analysis and multimodal tasks involving EEG or speech input [39]. The CNN component extracts frame-level features, while the LSTM captures

temporal evolution, enabling more nuanced emotion interpretation that accounts for both static and dynamic cues.

Transformer-CNN hybrids also show promise by merging local spatial priors from CNNs with the global attention capabilities of Transformers [40–43]. These models enhance robustness, especially in environments with high variability or limited annotated data. Ensemble approaches further strengthen classification performance by combining diverse learners to reduce variance and improve recognition of minority emotion classes [23, 44]. Collectively, these strategies demonstrate the increasing importance of integrated approaches in achieving reliable emotion recognition across multiple domains.

VI. Evolution Timeline

More recently, the field has increasingly shifted toward hybrid and multimodal architectures that combine convolutional networks, recurrent networks, transformers, and attention mechanisms. These integrated approaches seek to exploit complementary spatial, temporal, and contextual representations to improve robustness and generalization across heterogeneous environments [39–43].

This architectural evolution, as demonstrated in Fig. 2, is a clear progression from single-modality feature extraction toward multimodal affective intelligence systems capable of supporting increasingly complex real-world emotion recognition applications.

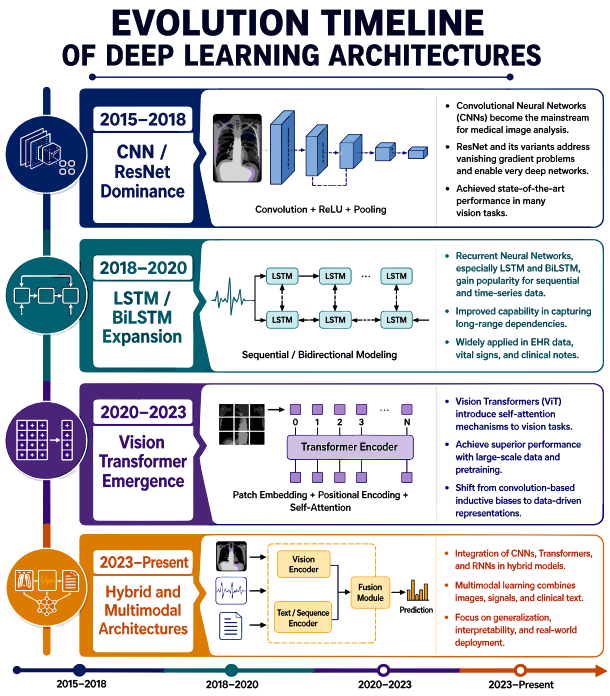

Fig. 2. Evolution of deep learning architectures for emotion recognition from 2015 to the present.

VII. Bias and Fairness Considerations

Emotion recognition systems are highly sensitive to dataset bias, which can arise from imbalanced demographic representation, lighting variations, occlusion, and cultural differences. High performance on curated datasets, such as CK+ [32], does not guarantee generalization, as demonstrated by the lower accuracy of models on datasets like FERPlus and AffectNet, which display emotions in more naturalistic and culturally diverse contexts [6, 41]. This discrepancy highlights the risk of overestimating model performance when evaluations rely solely on controlled datasets.

Research addressing these limitations emphasizes approaches such as data augmentation [41], adversarial training for robustness under noisy annotations [37, 38], and occlusion-aware modeling for masked-face scenarios [45]. These methods help mitigate representational bias and enhance fairness across demographic groups. Addressing bias is particularly critical in real-world deployments such as healthcare, education, and public safety, where misclassification could have social or ethical consequences.

A. Bias Quantification and Fairness Enhancement

Although overall accuracy remains the most commonly reported evaluation metric, increasing attention has been directed toward fairness and demographic generalization in emotion recognition systems. Differences in age, gender, ethnicity, and cultural background may influence both emotional expression patterns and model prediction behavior. Consequently, evaluating model performance across demographic subgroups has become increasingly important for ensuring equitable deployment in real-world environments.

Bias mitigation strategies may be implemented at multiple levels. At the dataset level, balanced sampling, demographic diversification, and synthetic data augmentation can improve representation across underrepresented groups. At the model level, adversarial debiasing techniques and fairness-aware optimization approaches can reduce demographic dependence in learned representations. At the evaluation level, metrics such as subgroup accuracy, demographic parity, and equalized odds can complement traditional performance measures by providing deeper insight into model behavior across diverse populations.

Addressing fairness remains particularly important in high-impact domains such as healthcare, education, and public services, where biased predictions may contribute to unequal outcomes among demographic groups.

Future evaluations should report subgroup-level performance metrics alongside overall accuracy to better assess fairness and demographic generalization.

VIII. Application Domains

Emotion recognition has numerous applications across various fields, including healthcare, education, IoT systems, and assistive technologies. In healthcare, CNN and hybrid models have been utilized to support diagnosis and monitor patient emotional states, particularly when combined with physiological data or contextual information [4, 28, 45].

Educational applications leverage lightweight models, such as MobileNet, to provide real-time insights into

student engagement and emotional responses on mobile and embedded platforms[12–14, 34], enabling instructors to adjust their teaching strategies dynamically.

In IoT environments, architecture selection must balance accuracy and resource constraints, favoring models optimized for speed and efficiency. Conversely, research and high-stakes applications benefit from the representational depth of more complex models such as ViTs and hybrid architectures. These diverse applications. demonstrate that architectural selection is not merely a matter of accuracy, but must also account for operational requirements, deployment constraints, and domain-specific challenges.

IX. Experimental Methodology

A. Research Framework

This study investigates the feasibility of detecting clinically observable affective states using a hybrid deep learning framework designed for temporally evolving facial behavior in elderly palliative care patients. Unlike traditional emotion recognition systems that infer psychological emotion categories, the proposed framework models observable behavioral manifestations aligned with caregiver assessment practice.

The experimental methodology operationalizes hybrid deep learning concepts discussed earlier in this review through a structured implementation pipeline consisting of dataset construction, spatial–temporal modeling, transfer learning, and cross-patient evaluation.

B. Clinical Dataset Acquisition and Annotation

Facial video recordings were obtained from institutional palliative care environments under approved ethical protocols. Data collection emphasized minimal intrusion to preserve patient comfort and ecological validity.

1) Behavioral state definition

Instead of categorical emotional labels, annotations represent caregiver-observable patient conditions:

- Neutral
- Discomfort
- Pain
- Distress
- Fatigue
- Relief

This formulation reframes affect recognition as condition monitoring, reflecting real clinical interpretation rather than emotional inference.

2) Temporal dataset construction

The clinical video recordings were preprocessed using a fixed temporal sampling strategy to generate consistent input sequences for behavioral state modeling. Table IV summarizes the temporal dataset construction parameters used for frame extraction, window segmentation, and behavioral labeling.

Temporal segmentation transforms the learning problem from static image classification into behavioral progression modeling.

TABLE IV. Temporal Dataset Construction Parameters

Parameter	Setting
Frame rate	5 fps
Window duration	~2 s
Label strategy	Majority interval labeling

C. Hybrid Model Architecture

To capture complementary representations, a hybrid feature-level architecture was implemented that integrates convolutional and transformer-based encoders.

1) Spatial feature encoders

Table V shows the functional contribution of the four representative architectures selected.

Each backbone generates latent embeddings rather than independent predictions.

TABLE V. Functional Contribution

Model	Functional Contribution
CNN	Local facial structure extraction
ResNet	Deep hierarchical representation
MobileNetV2	Efficient computation
Vision Transformer	Global dependency modeling

2) Feature fusion mechanism

Embeddings from all architectures are concatenated into a composite feature vector, which is then processed by a multilayer perceptron classifier. Feature-level fusion enables interaction among heterogeneous representations and improves learning stability compared with decision-level ensemble approaches.

3) Temporal modeling

Sequential modeling captures sustained behavioral evolution across time windows. This allows detection of persistent patient conditions rather than transient facial appearances. Temporal learning is critical in clinical populations where expressive intensity is typically weak or prolonged.

D. Training Strategy and Transfer Learning

Training followed a three-stage protocol:

Stage 1—Pretraining: General facial representations learned using FER2013.

Stage 2—Domain Adaptation: Fine-tuning using temporally annotated clinical data.

Stage 3—Evaluation: Testing on unseen patient sequences.

The hybrid deep learning framework was trained using a standardized configuration for model optimization and behavioral state classification. Table VI summarizes the training framework, optimization algorithm, loss function, training strategy, and computational environment employed during model development.

As shown in Table VI, the proposed hybrid framework was implemented using the PyTorch deep learning framework and optimized with the Adam optimizer. Categorical Cross-Entropy was employed as the loss function because the model performs multiclass behavioral state classification. Mini-batch learning was adopted to improve training efficiency and optimization stability,

while all experiments were conducted on a cloud-based GPU to accelerate model training.

TABLE VI. TRAINING CONFIGURATION

Parameter	Setting
Framework	PyTorch
Optimizer	Adam
Loss	Categorical Cross-Entropy
Training Mode	Mini-Batch Learning
Hardware	Cloud-based GPU

E. Evaluation Protocol

To prevent subject leakage, Leave-One-Video-Out (LOOV) validation was employed. Each fold evaluates model performance on an unseen patient, ensuring that reported accuracy reflects generalization capability rather than memorization. This evaluation paradigm approximates real clinical deployment conditions.

X. RESULTS AND DISCUSSION

The experimental findings operationalize the hybrid architecture analysis presented in earlier sections by validating its applicability under realistic clinical observation conditions.

A. Experimental Evaluation Using Leave-One-Video-Out Validation

To evaluate the real-world applicability of the proposed hybrid deep learning framework, model performance was assessed using the Leave-One-Video-Out (LOOV) validation method.

Under this protocol, the model was trained on all available patient recordings except one, with the excluded video serving as an independent test sample. This evaluation strategy simulates realistic clinical deployment conditions in which the monitoring system encounters previously unseen patients rather than familiar facial data.

Each validation fold, therefore, represents cross-patient generalization, preventing frame-level leakage and ensuring that performance reflects behavioral learning rather than subject memorization.

B. Performance Interpretation and Generalization

Across five validation folds, consistent convergence behavior was observed during training. Loss values decreased steadily across epochs, indicating stable optimization of the learned spatial–temporal representations derived from facial behavioral sequences.

The LOOV performance summary yielded:

- Mean Accuracy: 0.49 ± 0.45
- Mean Macro-F1: 0.0875 ± 0.075

Although numerical accuracy appears moderate relative to conventional emotion recognition benchmarks, this outcome must be interpreted within the clinical evaluation context. High variance across folds indicates substantial behavioral differences among patients, a characteristic intrinsic to elderly palliative populations rather than a modeling deficiency.

The relatively low Macro-F1 score is primarily attributable to severe class imbalance within the clinical dataset rather than unstable model learning. Minority behavioral states such as pain, distress, fatigue, and relief were substantially underrepresented compared with neutral conditions, limiting the model's ability to learn equally discriminative representations across all classes. This interpretation is further supported by the confusion matrix analysis, which demonstrates systematic prediction behavior rather than random classification errors.

Some unseen patient videos were successfully classified, while others led to prediction failure. This behavior suggests that the learned model captures adaptive behavioral recognition patterns, but full cross-patient invariance remains challenging due to individual physiological and expressive variability.

C. Dataset Imbalance and Behavioral Distribution Analysis

Prior to interpreting classification outcomes, the distribution of behavioral states was examined.

The visualization in Fig. 3 shows a strong dominance of neutral behavioral states in the dataset. Minority clinical conditions—including pain, distress, fatigue, and relief—appear substantially underrepresented.

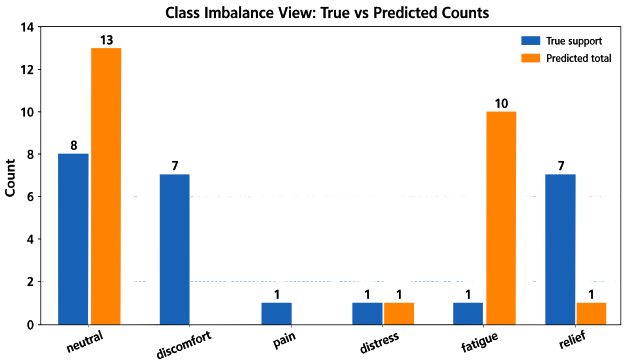

Fig. 3. Class imbalance view.

This imbalance produces two observable learning effects:

- Dominant behavioral states are over-predicted.
- Minority clinical states receive limited representation during training.

Consequently, overall accuracy remains moderate while Macro-F1 scores decrease, reflecting unequal class learning rather than unstable model behavior.

From a clinical standpoint, this outcome mirrors real hospital environments, where patients remain in stable or resting conditions for extended periods and expressive distress events occur less frequently.

The confusion structure demonstrates systematic prediction tendencies rather than random classification errors, revealing meaningful behavioral encoding by the hybrid model.

D. Confusion Matrix Analysis and Behavioral Interpretation

To further analyze prediction behavior, pooled LOOV results were examined using a confusion matrix (Fig. 4).

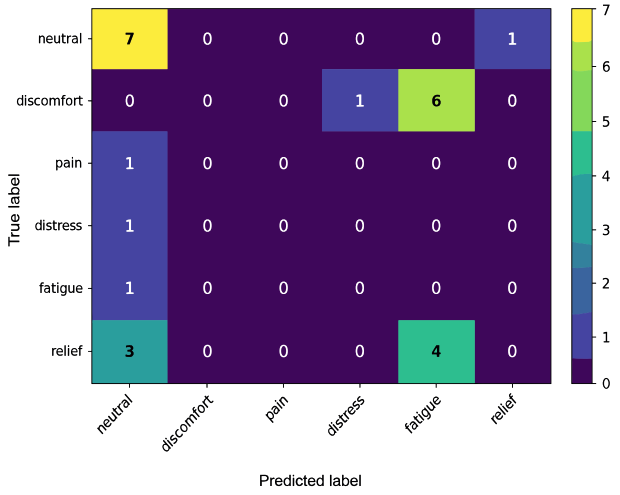

Fig. 4. Confusion matrix (pooled LOOV predictions).

1) Reliable neutral state detection

Neutral states achieved the highest recall (≈0.88), indicating strong system capability in identifying physiologically stable patient conditions.

This result confirms effective learning of baseline behavioral patterns.

2) Collapse toward neutral predictions

Pain, distress, fatigue, and relief were frequently classified as neutral. Rather than representing algorithmic failure, this tendency reflects the subtle manifestation of clinical affective states, which often lack strong facial activation.

The system therefore prioritizes detection of behavioral calmness versus deviation intensity.

3) Discomfort-fatigue overlap

Discomfort samples were commonly confused with fatigue or distress. Clinically, these conditions share observable characteristics such as:

- reduced facial movement,
- partial eye closure,
- decreased motor activity.

The model thus appears to encode behavioral intensity gradients instead of rigid emotional categories.

4) Relief interpreted as relaxed neutrality

Relief states were frequently predicted as neutral. In palliative care contexts, relief manifests as relaxation rather than expressive emotion, making such predictions clinically interpretable.

This observation strengthens the argument that observable behavioral modeling aligns more closely with caregiver perception than categorical emotion labeling.

E. Integrated Interpretation of Model Performance

Combining LOOV statistics, imbalance analysis, and confusion patterns reveals the operational characteristics of the proposed hybrid framework.

1) Correctly learned capabilities

The proposed hybrid framework successfully distinguishes between calm and expressive patient conditions, demonstrates sensitivity to variations in behavioral intensity, and adapts to patient-specific affective presentations. These findings indicate that the model is capable of learning clinically meaningful behavioral patterns despite the inherent variability observed across individual patients.

2) Current technical constraints

The proposed framework exhibits several limitations, including reduced separability among negative clinical states, sensitivity to dataset imbalance, and incomplete cross-patient generalization. These challenges are consistent with the complexity of clinical behavioral monitoring, where subtle affective cues, class imbalance, and patient-specific variability can limit classification performance.

F. Clinical Implications of Findings

The experimental outcomes indicate that affect recognition in palliative environments is more appropriately conceptualized as behavioral condition awareness rather than strict emotional classification.

The hybrid CNN-temporal learning framework therefore functions as a clinical state awareness indicator, capable of supporting observation of patient comfort and potential discomfort conditions even when categorical differentiation remains difficult.

Notably, classification errors follow clinically meaningful relationships rather than arbitrary misclassification patterns. This suggests that the system captures observable behavioral logic consistent with real caregiver assessment processes. Accordingly, the proposed framework supports continuous, non-intrusive monitoring intended to augment—not replace—professional clinical judgment.

XI. Emerging Research Directions

Several emerging technologies are expected to shape the future of emotion recognition research. Few-shot learning seeks to reduce dependence on large annotated datasets by enabling models to learn effectively from a limited number of training examples. Similarly, zero-shot learning aims to recognize previously unseen emotional categories by leveraging semantic relationships between known and unknown classes.

Federated learning has also emerged as a promising paradigm for privacy-sensitive applications, particularly in healthcare and educational environments. Rather than transferring sensitive data to centralized servers, federated approaches enable collaborative model training while preserving local data ownership and privacy.

Another rapidly growing research area involves cross-modal attention mechanisms that improve multimodal fusion by dynamically learning relationships between facial expressions, speech signals, textual content, and physiological measurements. These techniques provide greater robustness when individual modalities become noisy or unavailable.

Finally, researchers are increasingly transitioning from discrete emotion classification toward continuous affect modeling using valence-arousal representations. This

paradigm better reflects the dynamic and evolving nature of human emotions and may provide more meaningful insights for real-world affective computing applications.

Another promising direction involves hierarchical transformer architectures such as the Swin Transformer. Unlike conventional Vision Transformers that process fixed image patches globally, Swin Transformers employ shifted-window attention mechanisms that improve computational efficiency while preserving the ability to capture both local and global contextual information. This hierarchical design has demonstrated strong performance in computer vision tasks and offers significant potential for emotion recognition applications requiring high accuracy under computational constraints. Future studies may explore the integration of Swin Transformer architectures with multimodal emotion recognition frameworks to improve scalability and robustness in real-world environments.

XII. Conclusion

This paper presented a comprehensive comparative review of major deep learning architectures for emotion recognition, including Convolutional Neural Networks (CNNs), ResNet, MobileNet, Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Vision Transformers (ViTs). The analysis demonstrated that no single architecture consistently outperforms all others across every dataset, modality, and deployment environment. Instead, model effectiveness depends heavily on dataset characteristics, computational constraints, temporal requirements, and application objectives.

CNN and ResNet architectures continue to provide strong performance for visual emotion recognition tasks involving structured image datasets. MobileNet offers an effective balance between accuracy and computational efficiency, making it particularly suitable for real-time and resource-constrained deployments. LSTM and BiLSTM networks remain valuable for modeling temporal affective patterns in speech, EEG, and multimodal applications. Vision Transformers provide powerful global attention capabilities and show significant potential for large-scale affective computing research. Hybrid architectures further improve performance by combining complementary spatial and temporal representations.

The experimental evaluation conducted in this study further validated the practical applicability of hybrid deep learning frameworks for detecting clinically observable affective states in palliative care environments. Although challenges related to dataset imbalance, patient variability, and cross-subject generalization remain, the results demonstrate that hybrid architectures can effectively support behavioral condition awareness through continuous and non-intrusive monitoring.

Overall, the findings emphasize that successful emotion recognition requires context-driven architecture selection rather than reliance on a single model family. Future advances will likely emerge from multimodal integration, fairness-aware learning, and adaptive architectures capable of operating robustly across diverse populations and real-world environments.

XIII. Future Work

Future research should address several emerging challenges and opportunities in emotion recognition.

In the short term (1–2 years), efforts should focus on improving robustness under challenging conditions, including low-light environments, facial occlusion, mask usage, and limited training data. Few-shot learning and advanced augmentation techniques may help improve performance in data-scarce scenarios.

In the medium term (2–3 years), research is expected to emphasize multimodal fusion, fairness-aware learning, and privacy-preserving frameworks. Improved integration of visual, audio, textual, and physiological modalities may enhance recognition accuracy while reducing modality-specific limitations.

In the longer term (3–5 years), the field is likely to transition toward continuous affect modeling, real-time valence-arousal regression, and adaptive emotion understanding systems capable of operating across diverse populations and environments. Integration with emerging technologies such as brain-computer interfaces and advanced human-centered AI systems may further expand the practical impact of emotion recognition in healthcare, education, and intelligent interactive systems.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Armida P. Salazar conceived the study, conducted the literature review, designed and implemented the experimental methodology, performed the experiments, analyzed the results, and drafted the manuscript. Vladimir Y. Mariano provided research supervision, contributed to the study design and interpretation of results, critically revised the manuscript, and provided technical guidance throughout the study. Both authors reviewed and approved the final manuscript.

Acknowledgment

I would like to express my sincere gratitude to Dr. Eric Blancaflor for his invaluable support and insightful suggestions on both the structure and content of this paper. The author also extends sincere appreciation to National University for providing the academic environment and resources that made this research possible.

References

[1] H. Yang, J. Han, and K. Min, "EEG-based estimation on the reduction of negative emotions for illustrated surgical images," *Sensors*, vol. 20, no. 24, 7103, Dec. 2020. doi: 10.3390/s20247103

[2] W. Wei, Q. Jia, Y. Feng, and G. Chen, "Emotion recognition based on weighted fusion strategy of multichannel physiological signals," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–9, Jul. 2018. doi: 10.1155/2018/5296523

[3] W. Hua, F. Dai, L. Huang, J. Xiong, and G. Gui, "HERO: Human emotions recognition for realizing intelligent internet of things," *IEEE Access*, vol. 7, pp. 24321–24332, 2019. doi: 10.1109/ACCESS.2019.2900231

[4] T. Ichimura and S. Kamada, "A distillation learning model of adaptive structural deep belief network for affectnet: Facial expression image database," in *Proc. 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2020, pp. 450–455. doi: 10.1109/IIAI-AAI50415.2020.00096

[5] X. Huang, G. Zhao, M. Pietikäinen, and W. Zheng, "Expression recognition in videos using a weighted component-based feature descriptor," in *Proc. Scandinavian Conference on Image Analysis*, 2011, pp. 569–578. doi: 10.1007/978-3-642-21237-7_53

[6] A. Rehman, M. Mujahid, A. Elyassih *et al.*, "Comprehensive review and analysis on f acial emotion recognition: Performance insights into deep and traditional learning with current updates and challenges," *Computers, Materials & Continua*, vol. 82, no. 1, pp. 41–72, 2025. doi: 10.32604/cmc.2024.058036

[7] F. M. Talaat, Z. H. Ali, R. R. Mostafa, and N. El-Rashidy, "Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children," *Soft comput*ing, vol. 28, no. 9, pp. 6695–6708, 2024. doi: 10.1007/s00500-023-09477-y

[8] H. Jiang, D. Wu, R. Jiao, and Z. Wang, "Analytical comparison of two emotion classification models based on convolutional neural networks," *Complexity*, vol. 2021, no. 1, 2021. doi: 10.1155/2021/6625141

[9] P. Shukla and M. Kumar, "Explicating resnet for facial expression recognition," *International Journal of Computing, Intelligent and Communication Technology*, vol. 11, no. 3, 2022. doi: 10.57061/ijcict.v11i3.3

[10] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. British Machine Vision Conference*, 2016, no. 87, pp. 1–12. doi: 10.5244/C.30.87

[11] R. Kumari and J. Wasim, "A deep learning approach for human facial expression recognition using residual network—101," *Journal of Current Science and Technology*, vol. 13, no. 3, pp. 517–532, 2023. doi: 10.59796/jcst.V13N3.2023.2152

[12] Y. Miao, H. Dong, J. M. A. Jaam, and A. E. Saddik, "A deep learning system for recognizing facial expression in real-time," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 2, pp. 1–20, 2019. doi: 10.1145/3311747

[13] O. O. Arsirii and D. V. Petrosiuk, "Pseudo-labeling of transfer learning convolutional neural network data for human facial emotion recognition," *Herald of Advanced Information Technology*, vol. 6, no. 3, pp. 203–214, 2023. doi: 10.15276/hait.06.2023.13

[14] C. Wang, "Emotion recognition of college students' online learning engagement based on deep learning," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 17, no. 06, pp. 110–122, 2022. doi: 10.3991/ijet.v17i06.30019

[15] D. A. Dharma and A. Zahra, "Emotion recognition on multimodal with deep learning and ensemble," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, 2022. doi: 10.14569/IJACSA.2022.0131278

[16] N. Chen, "Visual recognition and prediction analysis of China's real estate index and stock trend based on CNN-LSTM algorithm optimized by neural networks," *Plos One*, vol. 18, no. 2, e0282159, 2023. doi: 10.1371/journal.pone.0282159

[17] R. Pereira *et al*., "Systematic review of emotion detection with computer vision and deep learning," *Sensors*, vol. 24, no. 11, 3484, 2024. doi: 10.3390/s24113484

[18] M. Chen, "Emotion analysis based on deep learning with application to research on development of Western culture," *Frontiers in Psychology*, vol. 13, 2022. doi: 10.3389/fpsyg.2022.911686

[19] M. Kastrati, M. Biba, A. S. Imran, and Z. Kastrati, "Sentiment polarity and emotion detection from tweets using distant supervision and deep learning models," in *Proc. International Symposium on Methodologies for Intelligent Systems*, 2022, pp. 13–23. doi: 10.1007/978-3-031-16564-1_2

[20] W. H. Hwang, D. H. Kang, and D. H. Kim, "Brain lateralisation feature extraction and ant colony optimisation-bidirectional LSTM network model for emotion recognition," *IET Signal Processing*, vol. 16, no. 1, pp. 45–61, 2022. doi: 10.1049/sil2.12076

[21] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER: Facial emotion recognition with vision transformers," *Applied System Innovation*, vol. 5, no. 4, 2022. doi: 10.3390/asi5040080

[22] S. A. Hasanah, A. A. Pravitasari, A. S. Abdullah *et al.*, "A Deep learning review of ResNet architecture for lung disease identification in CXR image," *Applied Sciences*, vol. 13, no. 24, 13111, 2023. doi: 10.3390/app132413111

[23] S. Khan, M. Naseer, M. Hayat *et al.*, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022. doi: 10.1145/3505244

[24] C. P. Vijay, "Literature review : Emotion-based music recommendation system with age detection," *Interantional Journal Of Scientific Research in Engineering and Management*, vol. 8, no. 4, pp. 1–5, 2024. doi: 10.55041/IJSREM31364

[25] I. A. Ganie, A. Gupta, and A. Oberoi, "A VGG16 based hybrid deep convolutional neural network based real-time video frame emotion detection system for affective human computer interaction," *International Journal for Research in Applied Science & Engineering Technology*, vol. 11, no. 11, pp. 1187–1195, 2023. doi: 10.22214/ijraset.2023.49221

[26] I. P. R. E. Wicaksana, G. R. Davinsi, M. A. Afriyanto *et al.*, "Systematic literature review: The influence and effectiveness of deep learning in image processing for emotion recognition," *Research Square*, 2024. doi: 10.21203/rs.3.rs-3856084/v1

[27] M. B. Abisado *et al.*, "An academic affect dataset: spontaneous facial expressions and head poses collected during online examination," in *Proc. 2020 the 4th International Conference on E-Society, E-Education and E-Technology*, 2020, pp. 83–87. doi: 10.1145/3421682.3421692

[28] Y. Xu, Y. S. Lin, X. Zhou, and X. Shan, "Utilizing emotion recognition technology to enhance user experience in real-time," *Computing and Artificial Intelligence*, vol. 2, no. 1, 1388, 2024. doi: 10.59400/cai.v2i1.1388

[29] S. Dwijayanti, M. Iqbal, and B. Y. Suprapto, "Real-time implementation of face recognition and emotion recognition in a humanoid robot using a convolutional neural network," *IEEE Access*, vol. 10, pp. 89876–89886, 2022. doi: 10.1109/ACCESS.2022.3200762

[30] H. Jiang, R. Jiao, Z. Wang *et al.*, "Construction and analysis of emotion computing model based on LSTM," *Complexity*, vol. 2021, no. 1, 2021. doi: 10.1155/2021/8897105

[31] M. Patil and H. Bodhe, "Movie and music recommendation system based on facial expressions," *International Journal for Research in Applied Science and Engineering Technology*, 2024. doi: 10.22214/ijraset.2024.58355

[32] N. Kumari, "Facial emotion detection using residual neural network," *International Journal of Applied Engineering & Technology*, vol. 4, no. 3, pp. 458–463, 2022. doi: 10.5281/zenodo.15979544

[33] Y. Chen and J. He, "Deep learning-based emotion detection," *Journal of Computer and Communications*, vol. 10, no. 2, pp. 57–71, 2022. doi: 10.36227/techrxiv.18866159

[34] H. Guo and Z. Gao, "Multimodal sentiment recognition based on Bi-LSTM and fusion mechanism," *Academic Journal of Computing & Information Science*, vol. 6, no. 6, 2023. doi: 10.25236/AJCIS.2023.060620

[35] S. G. Tejashwini and D. Aradhana, "Multimodal deep learning approach for real-time sentiment analysis in video streaming," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023. doi: 10.14569/IJACSA.2023.0140881

[36] A. A. M. Omer, "Image classification based on vision transformer," *Journal of Computer and Communications*, vol. 12, no. 4, pp. 49–59, 2024. doi: 10.4236/jcc.2024.124005

[37] S. A. Shanimol and J. Charles, "ResNet50 and GRU: A synergistic model for accurate facial emotion recognition," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 8, pp. 563–569, 2024. doi: 10.14569/IJACSA.2024.0150861

[38] A. Banerjee, P. Washington *et al.*, "Training and profiling a pediatric emotion recognition classifier on mobile devices," arXiv Preprint, arXiv:2108.11754, 2021.

[39] X. Qiu, "Improving robustness in emotion recognition via adversarial training," in *Proc. Fourth International Conference on Signal Processing and Machine Learning (CONF-SPML 2024)*, 2024, pp. 119–126. doi: 10.1117/12.3027127

[40] L. Song, M. Xia, L. Weng *et al.*, "Axial cross attention meets CNN: Bibranch fusion network for change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 21–32, 2022. doi: 10.1109/JSTARS.2022.3224081

[41] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in

the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019. doi: 10.1109/TAFFC.2017.2740923

[42] H. Wu *et al.*, "CvT: Introducing convolutions to vision transformers," in *Proc. the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31.

[43] P. Kulkarni and T. M. Rajesh, "A multi-model framework for grading of human emotion using cnn and computer vision," *International Journal of Computer Vision and Image Processing*, vol. 12, no. 1, pp. 1–21, 2021. doi: 10.4018/IJCVIP.2022010102

[44] C. Hewitt and H. Gunes, "CNN-based Facial affect analysis on mobile devices," arXiv Preprint, arXiv:1807.08775, 2018.

[45] M. Mukhiddinov, O. Djuraev, F. Akhmedov *et al.*, "Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people," *Sensors*, vol. 23, no. 3, 1080, 2023. doi: 10.3390/s23031080