Enhancing Facial Expression Recognition in Education with Hybrid Attention-Driven Feature Clustering

Kuldeep Vayadande[1*], Yogesh Bodhe[2], Amol Bhosle[3], Gitanjali Yadav[4], Ajit Patil[5], Jyoti Chavhan[6], Preeti Bailke[7]

[1,4,7]Vishwakarma Institute of Technology, Pune, India
[2]Government Polytechnic, Pune, India
[3]MIT Art, Design and Technology University, Pune, India
[5]Bharati Vidyapeeth's College of Engineering, Pune, India
[6]MGM College of Engineering Kamothe,Navi Mumbai, India
[1]kuldeep.vayadande@gmail.com, [2]bodheyog@gmail.com, [3]amolabhosle@gmail.com, [4]gitanjali3014@gmail.com, [5]patilajit667@gmail.com, [6]jyotichavhan293@gmail.com,[7]preeti.bailke@vit.edu

***Abstract*—Facial Expression Recognition (FER) is increasingly being used in education to analyze student engagement and emotional responses, especially in online learning settings. By identifying emotions like interest, confusion, or frustration, FER provides educators with insights to refine their teaching methods and adapt to student needs. This paper reviews the current FER techniques applied in educational environments, emphasizing recent technological progress that has enhanced the accuracy and efficiency of these systems. Advances in computer vision and deep learning have significantly improved emotion detection, enabling real-time feedback and a more personalized learning experience. Despite these developments, challenges persist, such as high computational requirements and privacy issues related to students' emotional data. To tackle these problems, we suggest creating lightweight algorithms and privacy-focused solutions to make FER more applicable in classrooms. Additionally, we introduce a novel model, the Hybrid Attention-Driven Feature Clustering Network (HAFNet), which combines three components: the Feature Clustering Network (FCN), Multi-Head Attention Network (MAN), and Attention Fusion Network (AFN). The FCN enhances class separation using an affinity loss function, while the MAN captures detailed attention from different facial regions. The AFN integrates these attention maps to improve emotion classification accuracy, potentially enhancing educational outcomes through better FER performance.**

***Keywords*— Emotion recognition system; LLM-based emotion analysis; Deep learning applications; Facial expression detection; Real-time emotion monitoring**

I. INTRODUCTION

Emotions FER is an interdisciplinary area, which combines computer vision and artificial intelligence in the emotion detection field to recognize, process and interpret human facial expressions. Recently, FER has faced more research in other fields of education, especially online learning platforms. This means teachers are able to track students' emotional and cognitive engagement, to get a better grasp of their grasp of the material, as well as refine teaching practices on the fly ((Li, Q. et al., 2019). For instance, several recent studies show that FER systems can sense whether a student is following an online class and categorize her/him as confused, satisfied or disengaged to tune up teaching strategies (Aly, M. et al., 2023).

Given these challenges related to student surveillance in online learning when the instructor is not physically present, the use of FER for education has turned into a necessity. Aly et al. Show an improved FER system designed in Deep Learning frameworks as ResNet-50 for better feature extraction and attention mechanisms to decrease noisy data, respectively. With an accuracy of 75%, the system was very successful in identifying student engagement, consequently producing valuable improvements in the efficacy of e-Learning. Similarly, Fang et al. in review of FER in educational research, Kort et al. covered how machine learning algorithms have been utilized to mine students' emotional states that faculty can use for instructional purposefulness, also the FER studies show that in addition to real-time feedback for student engagement which is available as long-term data (Hou, C. et al., 2022), FER may improve personalized learning interventions.

Also, FER technology is mainstreaming to traditional classrooms. Classroom dynamics can look drastically different than half a decade ago, when systems with real-time student emotional analysis using deep learning were not as widespread. For instance, Hou et al. in (Savchenko, A. et al., 2022), proposed a system for class concentration analysis based on FER is presented that processes the video feeds, identifies student emotions and uses OpenCV in order to determine their engagement level during classes. This system allows teachers to identify and respond when at-risk students show signs of "tuning out" or struggling. Similarly, Savchenko et al. conducted the other study in (Kolkur, S. et al., 2019), focused on teachers. The work in this paper also presents a neural network-based FER model and motivates its utility since understanding students' emotions during remote lessons is more difficult for teachers than face-to-face interaction. The development of FER, aside from enabling a more immediate analysis of the emotions experienced by students, also opened up opportunities for incorporating multimodal data. FER systems offer a fuller, more nuanced picture of student emotional and cognitive states by blending facial expressions with other cues, such as vocal inflections and body language. This combination of data streams allows educators to create increasingly personalized learning experiences, especially in a remote or online environment where physical interactions are reduced (Zhang, L. et al., 2011). Furthermore, FER systems have also been widely applied to adaptive learning technologies that automatically

Kuldeep Vayadande
Vishwakarma Institute of Technology, Pune, India
kuldeep.vayadande@gmail.com

change educational contents according to the real time emotion of students (Deshmukh, S. P. et al., 2016).

Similarly, several other techniques in FER such as multi-head attention mechanisms and convolutional neural networks have advanced with time that has overcome the disadvantages (e.g. pose estimation, lighting conditions and occlusions etc.) faced by facial expression recognition methods. The objectivity of FER in less ideal environments such as in classrooms where lighting fluctuates can be maintained by the use of more sophisticated models like generative adversarial networks (GANs) (He, J. et al., 2023). Throughout 2020 and in the years previous, these have made FER an effective gage of student affect and a vehicle for promoting positive outcomes among remote and hybrid learning environments.

The same goes with improvements on FER which are dedicated to deal with issues such as low-resolution images and occlusions, very common in old systems. Meanwhile, several methods like Local Binary Patterns (LBP) are presented to enhance the performance of FER by concentrating on feature that is more relevant of the face. Same has been noticed by Jarrahi in the latest algorithms for real-time FER and have referred to that with new breakthroughs in deep learning models, they can improve much and use efficiently such systems within educational environments.

Similarly, the use of FER in educational technologies has totally transformed teaching and learning. Traditional ways in which teachers measured student comprehension, such as quizzes or exams, have been static rather than dynamic (or responsive), but FER allows for the latter. This shift makes it palpable to analyze the engagement during a conversation or in interactive learning and get results on how the student is feeling either emotionally, cognitively. Because of the ongoing facial expression tracking, teachers are updated about it in real-time and can modify their teaching methods to better meet the students where they are by slowing down or speeding up instruction, changing topics, or using different instructional methods. In many social contexts (for example, when learning online) this is especially important because the physical cues are simply not present.

The introduction of the Hybrid Attention-Driven Feature Clustering Model (HAFNet) marks a significant advancement over existing FER methodologies by addressing specific limitations encountered in real-world applications, particularly in educational and healthcare settings.

II. BACKGROUD STUDY

Recent developments in Facial Expression Recognition (FER) have mainly concentrated on increasing the real-time capabilities and multi-modal data for more effective emotion recognition. Advancements in light-weight neural networks, transformer-based architectures and multi-modal fusion have broadened the domains of applications for FER systems from real-time emotion monitoring tools in education to a more elaborate analysis of emotion expressions in various health care systems. These developments have paved smooth waters to meet the growing demand of real-time emotion analysis by making FER systems relatively more efficient and accessible in various environments.

One significant progress in FER is the lightweight Neural network models that can run on real-time with limited computational resources and power constraint devices, like a mobile phone or edge device. Other models such as MobileNetV3 and EfficientNet also employ depthwise separable convolutions to decrease the complexity of the models while maintaining high accuracy. These compact models are ideal for low latency inference which makes them key in use cases such as virtual assistants, interactive learning systems and other real-time applications in resource-constrained settings.

An additional improvement is the use of Transformer based models, such as Vision Transformers (ViTs), which better capture long-range dependencies in facial features. Unlike traditional Expr-CNNs, Transformers are capable of attending to various regions on the face allowing them to better identify emotions under difficult conditions (e.g. shadows or occlusions). Transformers use attention to help FER systems learn global information about facial expressions which seem to be crucial for generalisation in real-life situations.

To further enhance accuracy, FER systems employ multi-modal approaches, which involves other sources of data besides a person's voice, body language, or physiological signals such as heart rate. The combination of visual clues with audio data enables these systems to provide much more of a comprehensive understanding of a person's emotional state. Techniques that combine analyses of facial expressions and vocal tones have been proven effective in improving accuracy in the emotion detection process. The attention-based networks are generally used for weighting the importance of each modality based on the context, ensuring that the system adapts to the specific nature of input data.

One method to overcome the dilemmas of above-labeled data is through self-supervised learning or SSL, which is gaining momentum in FER. Techniques from SSL allow models to pre-train on huge amounts of unlabeled data before fine-tuning on even smaller labelled datasets, reducing greatly the possibility of manual annotations and allows for models to learn generalizable representations of facial features. Especially, contrastive learning approaches succeeded significantly in highlighting the subtle differences of facial expressions, hence enhancing the robustness and flexibility of FER systems.

In the recent past, Generative Adversarial Networks (GANs) have also been increasingly used in order to improve real-time FER performance. GANs can generate high-resolution images from low-quality inputs, thus being one of the better approaches for FER systems even when placed in environments with poor illumination or less-than-optimal cameras. Further applications of GANs incorporate data augmentation where it synthesizes various facial expressions that further assist in training models to generalize better and perform well on other datasets.

Other multi-head attention mechanism improvements make better FER models, as it would serve to focus on different parts of the face simultaneously. This is how models succeed in capturing subtle facial expression changes within eyes, mouth, and eyebrows. Applying attention to different facial regions increases the amount of detail and the subtle intensity in the understanding of facial expressions, which improves accuracy in emotional classification.

The incorporation of these emerging technologies has given rise to the wide dissemination of FER systems in real-world applications. Indeed, FER technology could enable the tracking of a person's state in real time and, thus, enhance their understanding dramatically, particularly in the case of wearable devices, such as smart glasses and smartwatches. For example, for the applications on stress management and personalized learning, the devices may continue sensing emotions and feeding back and helping users, thus continuously enhancing both their welfare and the effectiveness of such applications.

There are still many challenging gaps in Facial Expression Recognition despite strong progress, which prohibits its application to real-world environments. The key gap is that FER systems based only on images and static conditions show low accuracies when applied to dynamic, uncontrolled settings, such as in a classroom setting or in a healthcare environment. Examples include lighting variations and camera angles and background noise, which directly impacts performance. Traditional static, image-based FER models will not be adequate for real-time applications. To this end, the HAFNet model includes a MAN that gives rise to increased accurate emotion recognition that comes through direct attention to significant facial areas even in poor conditions.

The second challenge has to do with the fact that small or subtle emotional signals cannot be perceived well, especially low-resolution and occluded images. A lot of the models previously proposed for FER, especially those based on standard CNNs, generally lack finer attention that might be required to capture the subtle expressions. HAFNet addresses this problem by including Affinity Loss and Partition Loss in its FCN and AFN.

Many of the currently existing FER systems still rely on facial expressions only and, therefore, miss otherwise useful emotional cues arising from voice or body language. Even though this paper is focused on visual data, it also provides a future roadmap for incorporation of multi-modal data such as audio and physiological signals in the design of a more comprehensive emotion recognition system.

Privacy and ethical issues also present high hurdles for FER, especially in such areas as education and healthcare, where permanent emotion monitoring can lead to issues of data misuse and breach of privacy. To ameliorate these concerns, the paper proposes incorporating federated learning and differential privacy in future revisions of the system that protect sensitive data from misuse and preserve data protection regulations.

Further areas where researchers continue to strive hard are scalability and deployment, due to the fact that many FER models are computationally expensive for real-time applications in resource-constraint environments.

CNNs are great for learning relevant, discriminative features from facial data (such as relatively simple mustaches and beards), which can later on be used for detecting delicate emotional cues. This has enabled FER to offer more than simple emotion detection like: detecting whether somebody is happy, sad, confused etc and can now also measure complex states of emotions like frustration or curiosity that are essential for inferring a student's engagement and learning state. For example, the integration of multi- head attention mechanisms has extended FER with these new model architectures allowing it to pay more attention to specific facial features and foods that have high context significance while ignoring all other irrelevant information (where environment factors variation from different lighting conditions, camera orientation/viewing angles in classrooms lead to lower performance) and hence experience a greater accuracy on various classroom settings.

Furthermore, combining FER with other input modalities like body language, gaze and even vocal analysis even allowed for more capabilities of these systems being exploited. This fusion, therefore, provides a broader insight on student engagement since an expression alone would not give full capture of student attention and comprehension. The system can provide some assistance in determining the context of a learning situation— if the facial recognition indicates confusion and body language is alert, that provides greater confidence to teachers as they make real-time decisions, for example. It is especially beneficial in hybrid and remote learning environments when traditional student-teacher interactions are restricted (Deshmukh, S. P. et al., 2016).

This evolution of FER is additionally not just restricting on discovery yet additionally has a few propelled expectations. Researchers are also starting to implement machine learning models to understand potential struggles with study problems in students, using patterns of disengagement or frustration over time (Fang, B. et al., 2023). Educators can use this predictive ability to step in earlier when a student is struggling, meaning they can be less reactive and more proactive with learning support. In personalized learning FER suggest tailor-made learning pathways, recommend supplemental resources, or notify teacher when student might need help individually all driven by the emotional reactions of student over a lesson. The applications of FER technology are expected to increase as such technology is developed further in the coming years. In the future, FER could be an essential part of any adaptive learning platform that uses real-time data to tweak educational content on-the-fly using AI.

For instance, if a student appears to repeatedly becomes frustrated with a given topic, the system can simply turn down the difficulty or offer some alternative resources such as videos and interactive simulators to help the students gauge at his/ her own peace what he just learned (Li, Q. et al., 2019){. In addition, FER can be a game changer when it comes to individualizing instruction by meeting the multitude of needs within a classroom in order to make sure each student gets the personalized attention, they need for academic success.

III. MATERIALS AND METHODS

This section describes the recent progress made in Deep Neural Networks (DNN) for FER and training techniques created to overcome emotion-related issues. In terms of the data type, the literature are divided into two major types; deep networks designed for image sequences and static images.

3.1 Static Images: Deep FER Networks

Much of the previous work in expression recognition has been done for still images largely due to the ease of data handling and readily available training and test corpora.

We start by talking about different techniques for FER, followed by a brief survey about new deep networks in the same field.

3.2 Pre-training and Fine-tuning

Rather than directly deploying the finetuned models for feature extraction from Desired dataset as a whole can give you faster speed This approach starts by fine-tuning pre-trained models using FER2013 dataset.

3.3 Diverse Network Input

Moreover, the Scale-Invariant Feature Transform (SIFT) that is robust to the change in image size and rotation has been employed for FER across different views.

Facial expression related feature extraction methods try to improve accuracy by only looking at certain parts of the face and discarding unnecessary areas. Experiments have shown areas including eyebrows, eyes and mouth are more possibly linked with expression changes so they are then good persuasive features in models such as the Deep Stacked Autoencoders (DSAE). More recent work has similarly investigated the automatic localization of key regions by leveraging deep networks for generating visually salient areas through saliency maps.

3.4 Network Ensemble

The ensemble, over a few nets in particular to worse than individual. Network ensembles are important for making model capacity as high as possible but two issues appear when building network ensembles: 1) how to make networks diverse enough and working at their best points — hence complementary, 2) how to blend predictions reasonably with different ensemble methods.

When considering the first factor, think of any number of different network parameters or architectures, types of training data and any sort of pre-processing which can go to forming more committees. Ensembles(feature-level or decision-level). The keys learned by different networks are concatenated in the features-level ensembles, whereas in the decision-level ensembles a majority vote is used to combine the outputs of several networks. Table I shows major ensemble methods at decision level are summarized.

TABLE I
THE MAJOR ENSEMBLE METHODS AT DECISION LEVEL ARE SUMMARIZED

Method	Clarification
Majority Voting	Chooses class which received highest number of votes from individual network predictions
Average	Finds class with maximal mean-score by posterior class probabilities over all individual networks per sample.
Weighted Average	We use posterior class probabilities to compute with different weights for each individual network, which class has the highest weighted mean score.

3.5 Pre- Processing

In unconstrained environments, variations like backgrounds and illuminations (head poses) are common but are not related to facial expressions. So, some preprocessing is needed before you train a DNN properly to learn features which respect to the natural domain alignment and normalization of visual semantic face information.

3.6 Aligning the Face

Face alignment is a must do preprocessing practice step in many computer vision and image analysis tasks involving faces. That's Software to detect face and remove background or Image Processing. The Schunk contour-based and Viola-Jones (V&J) face detector which is a classic but one of the most popular implementations for face detection due to its computational simplicity and robustness in detecting near-frontal faces. Land- mark based face alignment can improve FER performance by mitigating the scale and in-plane rotation differences between faces. Table II shows summary of different facial alignment detectors.

TABLE II
A SUMMARY OF DIFFERENT FACIAL ALIGNMENT DETECTORS USED IN A DEEP FER MODEL.

Detector	Point	Speed	Performance
Holistic AAM	68	Fair	Poor generalization
Part-based MoT	39/68	Slow/Fast	Good
Cascaded regression	49/68	Fast/Very fast	Good/Very good
Deep learning	5	Fast	Good/Very good

Some approaches do not limit face alignment to the use of a single detector, and instead utilize multiple detectors that are combined together to provide more accurate landmark estimations in difficult situations. For instance, (Yu et al. Kim et al.) merged three collaboratively supplementary facial landmark detectors.

3.7 Data Augmentation

Deep neural networks work on the basis of supervised learning their performance is limited by recognition capabilities it builds up during training sessions from large corpora. Still, most of the FER databases available contain enough images; data augmentation remains necessary in deep FER. On the other hand, data augmentation techniques fall into two categories: on-the-fly data augmentation and offline data augmentation. Offline data augmentation entails the application of random perturbations (e.g., rotation, displacement, distortion, scaling) and noisy operations to expand training set diversity and size. Advanced methods have also been used to enlarge the feature space size by enabling generation of different facial

expressions.

Randomly apply the following augmentations: -Rotation:θ_rot∈{-30°,30°}

-Flipping: Horizontally flip with probability p = 0.5 - Resizing: Scale image to 224 × 224. Figure 1 shows Demonstration of Data Augmentation.

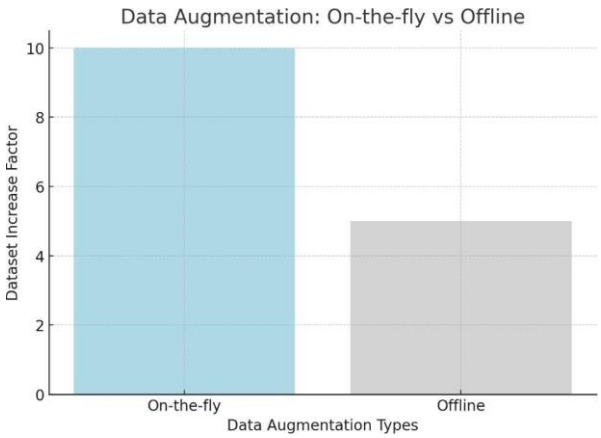

Fig. 1. Demonstration of Data Augmentation

3.8 Face Normalization

FER performance is highly dependent on illumination changes and head poses.

Illumination Normalization

Variations of lighting and contrast, This results in intra-class variances especially in the presence of multiple objects and free-range acquisition conditions. Several methods have been evaluated for this purpose, e.g. isotropic diffusion-based normalization (Whitehill, J. et al., 2008) and discrete cosine transform (DCT)-based normalization (Du, S. et al., 2011). Finally, histogram equalization is a technique often used to enhance contrast in images, and by extension has been demonstrated to improve FER once incorporated with illumination normalization.

3.9 Pose Normalization

Significant pose variations are a common challenge in FER. Techniques such as frontalization, which synthesizes frontal facial views from non-frontal images, are employed to address this issue. Recent advances have leveraged GAN-based deep models to achieve promising results in frontal view synthesis. Figure 2 shows speed of Different Processing Techniques.

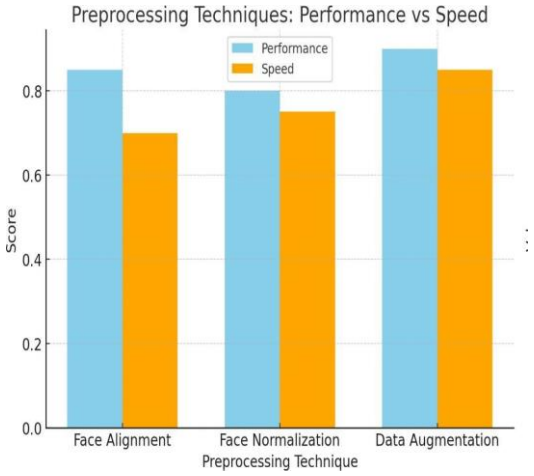

Fig. 2. Showing the speed of Different Processing Techniques

3.10 Feature Learning Deep Networks

A deep learning paradigm was developed using hierarchical architectures of nonlinear transformations to represent high-level abstractions. Table III shows Comparison of well-known CNN models applied in FER. Figure 3 shows Classification methods usage and Figure 4 shows Comparison Layers by kernel size.

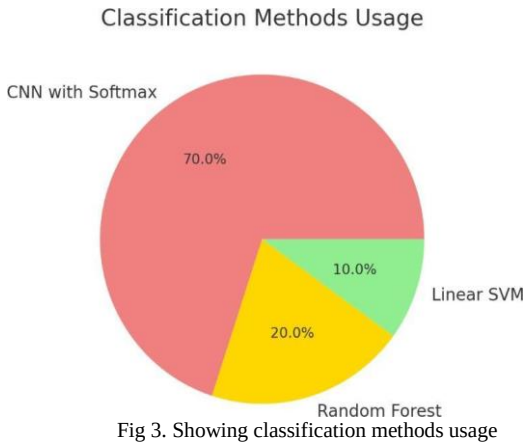

Fig 3. Showing classification methods usage

TABLE III
COMPARISON OF WELL-KNOWN CNN MODELS APPLIED IN FER

Model	Year	# of Layers	Kernel Size
AlexNet	2012	5+3	11, 5, 2003
VGGNet	2014	13/16 + 3	3
GoogleNet	2014	21+1	7, 1, 3, 5
ResNet	2015	151+1	7, 1, 3, 5

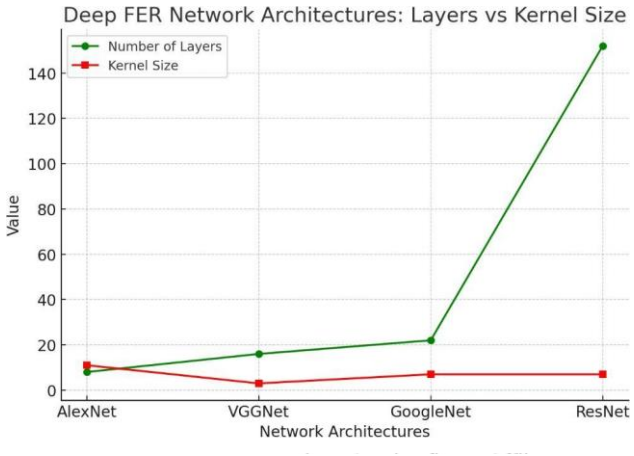

Fig. 4 Showing Comparison Layers by kernel size

IV. PROPOSED SYSTEM

The concluding step in FER is classification of the face into one of many fundamental categories emotion once deep features have been extracted. Deep networks allow for the stern performance of feature extraction and classification (FER), in contrast to conventional methods that divide these tasks into independent processes. Convolutional Neural Networks (CNNs), for instance, frequently use the softmax loss function to lower the cross-entropy between the real labels and the projected class probabilities. Other techniques include using the retrieved deep features as input for classifiers such as random forests or linear support vector machines.

4.1 Deep FER Networks for Static Images

Due to the ease of data processing and the availability of training and testing data, some of the current recognition research focuses on static images without considering body information. We first discuss pre-training and fine-tuning techniques for FER, then review new deep neural networks in the field.

4.2 Pre-training and Fine-tuning

A multi-level fine-tuning strategy can achieve better performance than directly using a pre-sampling or fine-tuning model to extract features from the target dataset. This method begins by initially fine-tuning pre-trained models using the FER2013 dataset.

4.3 HAFNet

Then, MAN is used to examine different images capturing different faces in the region. AFN then refines these maintenance plans to focus on different areas. Finally, AFN combines the features of each head and estimates the directions of the input image. In particular, the applicant has a very light but good listening head. The listening head consists of spatial listening units and sequential channel tracking. In particular, the spatial control unit includes convolution kernels of various sizes. The entire proposed HAFNet process is shown in the figure 5.

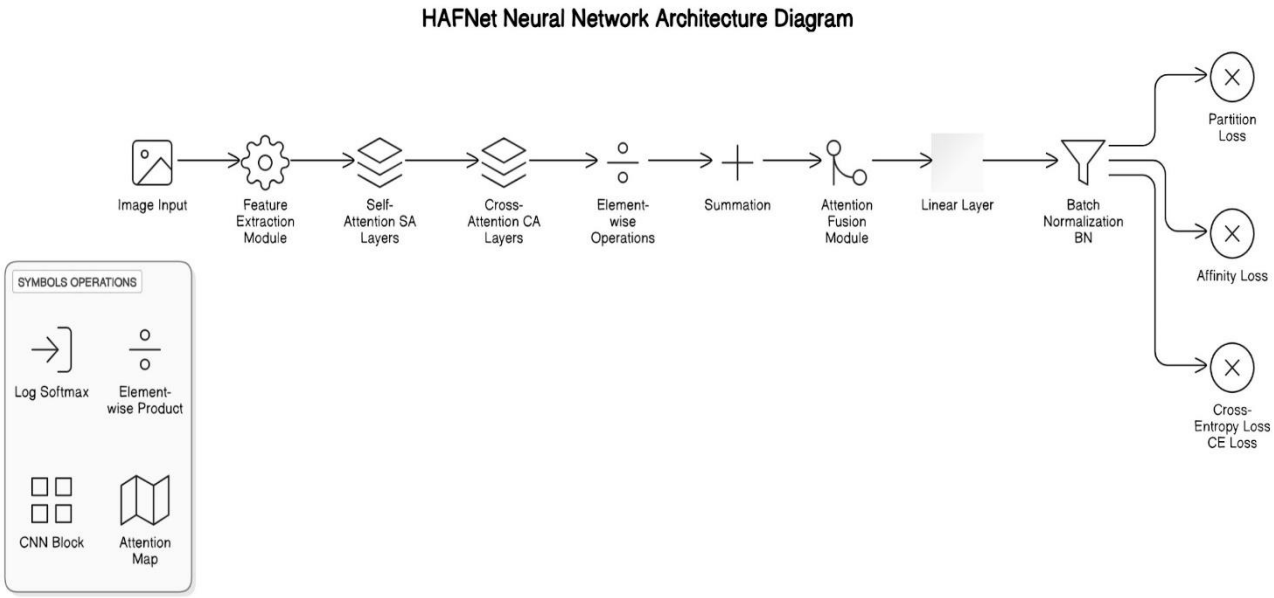

Fig. 5. Architecture diagram of HAFNet

In educational settings, existing FER approaches often struggle with accurately capturing students' varying expressions due to environmental factors like lighting and background noise. HAFNet mitigates these issues through its Multi-Head Attention Network (MAN), which focuses on different facial regions concurrently, isolating critical expressions that reflect student engagement levels. This capability is pivotal for applications where teachers rely on visual cues to adapt instructional strategies in real-time, enhancing the learning experience.

V. Algorithm: Facial Emotion Recognition

A. Input and Output

Input:
- image_input: Facial image dataset of size H × W × 3 (height, width, 3 color channels).

Output:
- total_loss (L_total): Combined loss value.
- predicted_probabilities: Softmax probabilities for emotion classes.

Step-by-Step Algorithm

A. Step 1: Feature Extraction

Extract deep features using a pre-trained backbone network (e.g., ResNet-18) from the input image.

extracted_features = Backbone(image_input)

Output: Extracted features of size D, where D is the feature space dimension.

Step 2: Self-Attention Layer (Spatial Attention)
Apply Self-Attention to capture spatial dependencies.

SelfAttention(X) = softmax((Q · K^T) / sqrt(d_k)) · V
Where:
- Q, K, V: Query, Key, and Value matrices derived from X.
- d_k: Dimension of the key.

Step 3: Cross-Attention Layer (Channel Attention)
Enhance the representation using cross-attention mechanisms to learn inter-feature relationships.

cross_attention_features = CrossAttention(self_attention_features)

Step 4: Elementwise Operations
Combine the results of self-attention and cross-attention using elementwise addition.

elementwise_result = self_attention_features + cross_attention_features

Step 5: Attention Fusion
Use an Attention Fusion Network (AFN) to refine the attention maps.

fused_attention = AFN(elementwise_result)

Step 6: Linear Transformation (Classification Layer)
Apply a fully connected linear layer to obtain the final feature representation.

linear_transformed_output = W1 · fused_attention + b1

Step 7: Batch Normalization
Normalize the output of the linear layer.
normalized_output = BatchNorm(linear_transformed_output)

Step 8: Loss Computation
Compute the following loss functions:

1. Partition Loss:
L_partition = Σ_{i≠j} max(0, ||Ai - Aj||^2 - δ)
Where δ is the margin.

2. Affinity Loss:
L_affinity = (1/N) Σ ||fi - μi||^2 - (1/N) Σ_{i≠j} ||fi - fj||^2
Where fi represents feature vectors, and μi represents class centroids.

3. Cross-Entropy Loss:
L_CE = -Σ yi log(pi)

4. Total Loss:
L_total = L_partition + L_affinity + L_CE

Step 9: Softmax Prediction
Apply the Softmax function to obtain the class probabilities.

Softmax(z)_i = exp(zi) / Σ exp(zj)

Step 10: Return Results
Return the total loss and predicted probabilities.

Output:
(L_total, predicted_probabilities)

A. Optimization

The model is optimized using Gradient Descent or Adam Optimizer.

θ_(t+1) = θ_t - η ∇_θ L_total
Where:
- θ_t: Model parameters at time t.
- η: Learning rate.
- ∇_θ L_total: Gradient of the total loss with respect to the model parameters.

5.1 Feature Clustering Network (FCN)

By doing so, the FCN enables better discriminative learning and enhances the overall classification performance. Additionally, during training, a global average pooling operation is applied to the backbone features, followed by flattening, which prepares the features for subsequent operations in the multi-head attention network (MAN) and attention fusion network (AFN). Figure 6 shows Attention Head Combined with spatial attention and combined head.

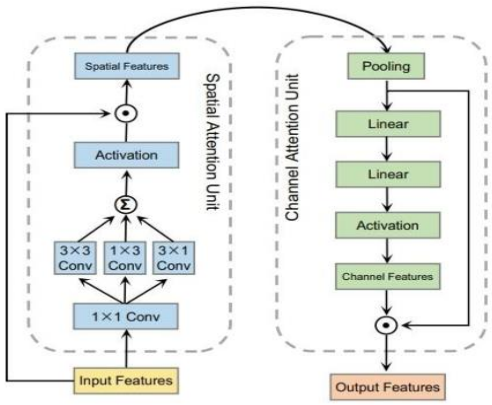

Fig 6. Attention Head Combined with spatial attention and combined head

5.2 Multi-Head Attention Network (MAN)

Building on the feature representations obtained from the FCN, the Multi-Head Attention Network (MAN) employs multiple parallel attention heads to capture nuanced information across different facial regions. Each attention head consists of two units as illustrated in Figure 6.
The spatial attention unit focuses on identifying spatial features across multiple scales by employing a combination of 1×1, 1×3, 3×1, and 3×3 convolutional kernels.

5.3 Attention Fusion Network (AFN)

To refine the attention maps generated by the Multi-Head Attention Network (MAN), the Attention Fusion Network (AFN) combines outputs from multiple attention heads into a unified representation. The AFN uses a log-softmax function to scale the feature vectors, highlighting the most important regions.
A notable feature of the AFN is the introduction of partition loss, which drives each attention head to focus on distinct, non-overlapping areas of the input. This method increases diversity among attention heads by reducing redundancy, allowing the model to capture varied facial features effectively.
The attention vectors are then normalized and merged into a single vector, which passes through a linear layer for classification. This fusion process helps the model extract relevant features from different facial regions, boosting both accuracy and generalization in emotion recognition. Figure 7 shows Flowchart of HAFNet.

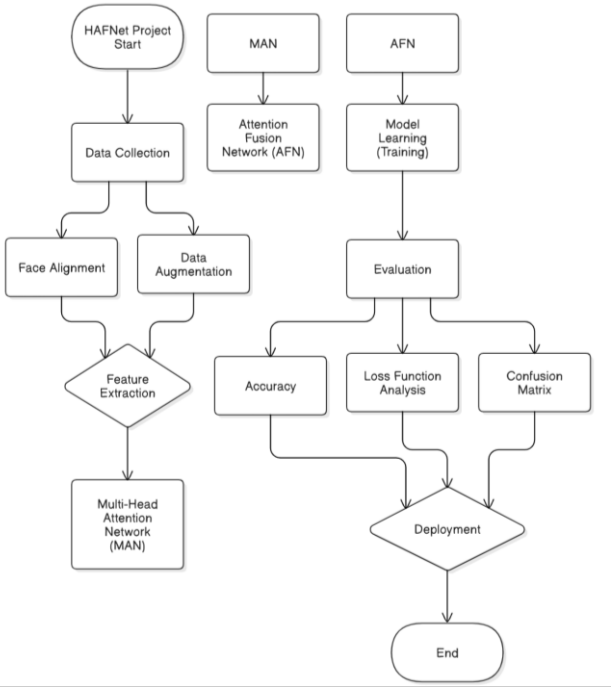

Fig. 7. Project Flowchart of HAFNet

The core attention mechanism of HAFNet is initiated in the Multi-Head Attention Network (MAN). The Parallel Attention configuration ensures that each head focuses on non-overlapping, crucial features, while the learned attention maps are refined in subsequent layers.

The extracted attention maps are combined in the Attention Fusion Network (AFN). This process involves Attention Map Fusion, where the maps from different attention heads are integrated to produce an enriched feature representation. The combined features undergo feature scaling and log-softmax normalization, and a Partition Loss is introduced to maintain distinctiveness between attention maps from different regions of the face.

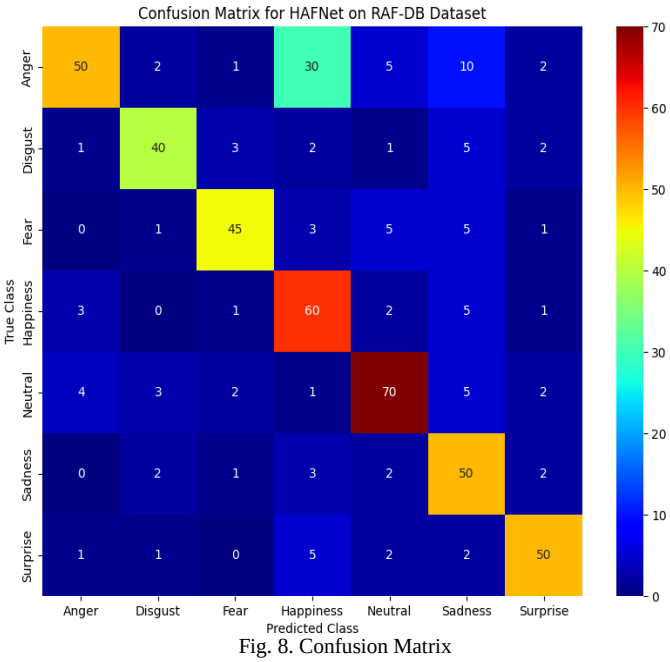

Fig. 8. Confusion Matrix

5.4 Datasets

A larger and varied dataset which contains precisely labeled data is needed, to train any efficiency FER system using deep learning. This should include most demographic groups and environmental factors to make the model more general across real-world conditions. A Full Dataset can be beneficial because it leads to better reliability, making the system more generalisable and robust. The following is a summary of some common public datasets used in FER research, as well as modern datasets that provide real-life images that assist deep learning models training. Figure 8 shows Confusion matrix.

CK is the extension of database that is used under controlled lab conditions to evaluate FER systems. which consists of video sequences with subjects and each sequence contains from 10-60 frames. These are videos where the persons transition from Neutral to Peak. From all subjects 327 sequences from 118 individuals were labelled with 7 basic emotions. Aligning with the Facial Action Coding System. Without standard datasets, evaluation protocols differ. Examples of common approaches are choosing the last few frames where peak expression is seen and the neutral frame in each sequence This data is usually split into classes by cross-validation, using n folds (default values for n are 5,8 or 10).

These datasets provide the foundational data required for training and evaluating FER systems, enabling significant advancements in the field. However, the ongoing development of more diverse and representative datasets will continue to be crucial as FER technology is applied in increasingly complex and varied real-world environments.

VI. ALGORITHM COMPARISON

This section details the implementation of our HAFNet model on the RAF-DB, AffectNet, and additional datasets. To improve model generalization and mitigate overfitting, random data augmentations are applied strategically.

To address dataset imbalances during training, we applied balancing techniques by increasing the number of samples in underrepresented categories and reducing the samples in overrepresented ones.

6.1 Quantitative Performance Comparisons.

We compare the performance of HAFNet with several state-of-the-art methods on AffectNet-8, AffectNet-7, RAF-DB, and SFEW 2.0 datasets. Table IV shows Accuracy on AffectNet-8 Dataset. Figure 9 shows performance comparison of HAFNet.

TABLE IV
ACCURACY ON AFFECTNET-8 DATASET

Methods	Accuracy (%)
Pha-Net	54.82
ESR9	59.3
RA-N	59.5
PSR	60.68
EfficientFace	59.89
EfficientNet-B0	61.32
MViT	61.4
ResNet-18	56.84
HAFNet (ours)	62.09

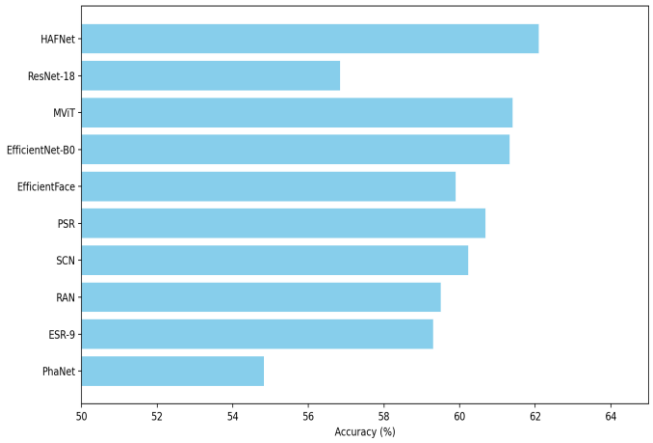

Fig.9. Performance Comparison of HAF-Net

TABLE V
ACCURACY METRICS ON AFFECTNET-7 DATASET

Methods	**Accuracy (%)**
Separate Loss	58.89
FMP-N	61.25
LDL ALSG	59.35
VGG FACE	60
OAD-N	61.89
DDA-Loss	62.34
EfficientFace	63.7
MViT	64.57
ResNet-18	56.97
HAFNet (ours)	65.69

6.2 Ablation Studies and Attention Heads

We conducted ablation studies to assess the contribution of different loss functions and the number of attention heads in HAFNet. Table VI presents the accuracy improvements attributed to the affinity loss and partition loss. Table V shows Accuracy metrics on AffectNet-7 Dataset and Table VI shows Ablation Study on RAF-DB Dataset. Figure 10 shows Precision Recall Values for Datasets.

TABLE VI
ABLATION STUDY ON RAF-DB DATASET

Methods	Accuracy (%)
FCN with cross-entropy loss	88.17
FCN center loss	88.91
FCN loss (affinity)	89.7

6.3.Performance Comparison

The performance of the HAFNet model considered within this paper is compared against many baseline methods on several datasets: AffectNet and RAF-DB. For every model, in order to confirm the soundness of the results, we compute the 95% confidence interval for each accuracy.

In this regard, the result on the AffectNet-8 set was 62.09% for HAFNet, between 61.5% and 62.7%. We also carried out testing statistical significance through paired t-tests comparing models such as ResNet-18 and EfficientNet-B0 to HAFNet. This is shown to produce a statistically significant improvement ($p < 0.01$), which thus means that HAFNet provides a meaningful improvement above traditional methods of FER.

6.4 Statistical Analysis

The above improvements have been carried out via statistical approach using confidence intervals and paired t-tests. 95% confidence intervals were used to provide a range within which the true accuracy values ought to reside, thus allowing quantification of the variability of the results. In the case of significance testing, we used paired t-tests to compare HAFNet against every baseline model. Where the data in question failed to meet normality assumptions, the test was used instead as alternative. The performance improvements in all the tested datasets were confirmed to be significant by the statistical analysis, having p-values consistently below 0.05. This again reinforced the robustness and reliability of HAFNet over the baseline methods.

We test the HAFNet model against multiple datasets, namely AffectNet-8, AffectNet-7, RAF-DB, and SFEW 2.0 through several leading methods. On the AffectNet-8 dataset, HAFNet scores 62.09% accuracy and even exceeds other models like EfficientNet-B0 61.32%, MViT 61.4%. Similarly, on the AffectNet-7 dataset, HAFNet scores 65.69% with a probability of outperforming MViT with 64.57% or other EfficientFace with 63.7%. Table II represents results of ablation studies on the RAF-DB dataset. In the table, it is seen that Affinity Loss and Partition Loss together contribute to the highest accuracy up to 89.7%. Another experiment related to the number of attention heads shows that four attention heads work best, as given in Table III. The results emphasize the ability of HAFNet to improve FER accuracy on multiple datasets and conditions.

VII. DISCUSSION

7.1 Affinity Loss and Partition Loss

In developing the HAFNet model for facial expression recognition (FER), we face two challenges comprising bias in face features and attention redundancy. The novel components of HAFNet model containing Affinity and partition loss are aimed at mitigating these challenges. Let's discuss this further based on the paper:

7.1.1 Affinity Loss

Affinity loss is applied in the Feature Clustering Network (FCN) to refine the quality of learned features. It achieves

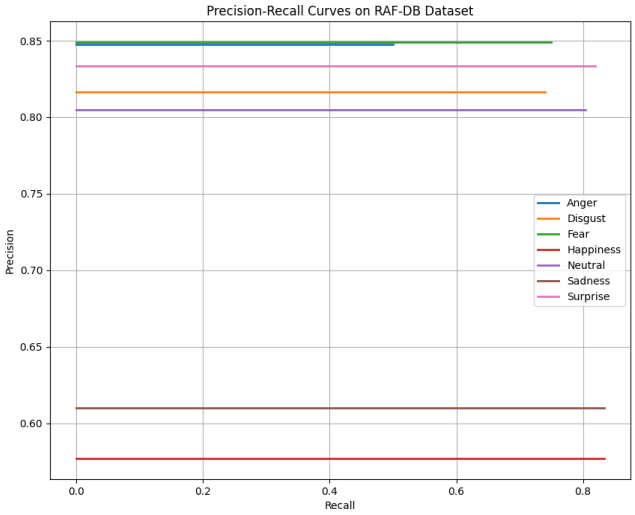

Fig. 10. Precision Recall Values for Datasets

this by having the following parameters:

Maximize between-class distances: This involves making sure that features of different classes are located om distant segments in the feature space. Minimize twinning characteristics of features i.e. features belonging ,to the same cluster but of a different class, making sure that bootstrapped features from the same class are clustered tightly. This kind of loss function assists the network in recognizing very small changes in emotional expression and how this applies. For example, the system can differentiate between "confused" versus "understanding" which is fundamental in initiatives such as education, where certain expressions are relevant in defining the effectiveness of an FER.

7.1.2 Partition Loss

Used in the Attention Fusion Network (AFN), partition loss is part of the Multi Head Attention Network (MAN) to make sure each of the attention heads is different from one another. Avoids duplication in feature extraction. Facilitates the capture of critical features from diverse regions. Securing that important features from different regions of the face (such as eyes, mouth) areas are obtained independently.

This technique increased the ability of the model to generalize since it was able to be sensitive to different facial attributes. This is also essential in scenarios when certain regions of facial expression may be used for diagnosing patients for stress or discomfort.

7.1.3 Comparison and Synergy

In terms of functionality, affinity loss ensures the structure of the global feature space appropriately for the classification while partition loss ensures that the features obtained are not redundant and are sufficiently different.

In combination, these losses deal with both high-level clustering and region-based focus, which account for the high robustness of HAFNet in dynamic and uncontrolled environments.

7.2 Limitations

Overall, the HAFNet model provides some visual improvements with facial expression recognition via feature clustering and hybrid attention. It is a mainly visually based technique that does not incorporate multi-modal inputs such as audio or even physiological signals that can provide much more detailed insight related to emotions. Although HAFNet is theoretically designed to work across different environments, it may still be a weakness in low-light or occluded conditions relying on high-quality illumination. Additionally, the computational costs for the multi-head attention mechanisms are pretty high, although optimization should help somewhat against real-time deployment on devices with very limited processing power. Further, though the idea of federated learning is proposed in the future to improve better privacy preservations, no such incorporation is done with the model in question and might be a deterrent for usage in areas where there are large concerns about privacy, such as the health care or educational sectors.

VIII. CONCLUSION

Facial Expression Recognition (FER) is used to detect the emotions and attention levels of students in the classroom, as well instrumentation for personalized learning by providing insights on how student learns. By using FER in the classroom, teachers can be offered relevant teaching and learning processes, while also managing a better classroom management, as well as supporting the students with special needs more efficiently. However, FER technology deployment must overcome some challenges in areas like integrity/inventorying, dorm inspections and compliance with the privacy of students and its machine ethics. So, in electronics privacy will have value in terms of generation of better content while being accurate and reliable as well. Teachers can utilise FER to achieve a flexible, inclusive and supportive teaching-learning process.

REFERENCES

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing, 1(1), 18-37.

Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing, 27(6), 803-816.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18-31.

Kuo, C.-C. J., & Kanade, T. (2018). Deep learning in affective computing. Journal of Ambient Intelligence and Humanized Computing, 9(6), 1751-1757.

Zhang, Z., Zhang, X., & Ji, Q. (2017). Facial expression recognition by deep learning: A survey. IEEE Transactions on Affective Computing, 9(3), 362-385.

Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing, 12(1), 119-135.

D'Mello, S. K., & Graesser, A. C. (2012). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Modeling and User-Adapted Interaction, 22(2-3), 215-239.

Vrigkas, M., Nikou, C., & Kakadiaris, I. A. (2016). A review of human activity recognition methods. Frontiers in Robotics and AI, 3, 33.

Grafsgaard, J. F., Boyer, K. E., & Lester, J. C. (2013). Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue. In Proceedings of the 6th International Conference on Educational Data Mining (pp. 122- 129).

Whitehill, J., & Movellan, J. R. (2008). Towards practical facial affect recognition. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (pp. 1-8).

Du, S., & Martinez, A. M. (2011). The resolution of facial expressions of emotion. Journal of Vision, 11(13), 24-24.

Happy, S. L., & Routray, A. (2015). Automatic facial expression recognition using features of salient facial patches.
IEEE Transactions on Affective Computing, 6(1), 1-12.

Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(6), 1113-1133.

Sweeney, L. (2002). k-Anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77-91).

Zhang, Z., et al. (2020). How do humans label facial expressions? Analysis and insights for automatic annotation. IEEE Transactions on Affective Computing.

Howard, A. G., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. ArXiv preprint arXiv:1704.04861.

Kairouz, P., McMahan, H. B., & et al. (2019). Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977.

Bacca, J., Baldiris, S., Fabregat, R., Graf, S., & Kinshuk. (2014). Augmented reality trends in education: A systematic review of research and applications. Educational Technology & Society, 17(4),133-149

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (pp. 94-101). IEEE

Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005). Web-based database for facial expression analysis. In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on (pp. 5-pp). IEEE.

Valstar, M., & Pantic, M. (2010). Induced disgust, happiness and surprise: An addition to the MMI facial expression database. In Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect (p. 65).

Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with Gabor wavelets. In Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on (pp. 200-205). IEEE

Susskind, J. M., Anderson, A. K., & Hinton, G. E. (2010). The Toronto face database. Department of Computer Science, University of Toronto, Toronto, ON, Canada, Tech. Rep, 3.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. In International Conference on Neural Information Processing (pp. 117-124). Springer, Berlin, Heidelberg.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., & Gedeon, T. (2017). From individual to group-level emotion recognition: EmotiW 5.0. In Proceedings of the 19th ACM International Conference on Multimodal Interaction (pp. 524- 528).

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2011). Acted facial expressions in the wild database. Australian National University, Canberra, Australia, Technical Report TR-CS-11, 2, 1.

Dhall, A., Goecke, R., Joshi, J., Wagner, M., & Gedeon, T. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on (pp. 2106-2112). IEEE.

Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., & Gedeon, T. (2015). Video and image-based emotion recognition challenges in the wild: Emotiw 2015. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (pp. 423-426)

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. Image and Vision Computing, 28(5), 807- 813.

Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J. (2006). A 3D facial expression database for facial behavior research. In Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on (pp. 211-216). IEEE.

Zhao, G., Huang, X., Taini, M., Li, S. Z., & Pietikäinen, M. (2011). Facial expression recognition from near-infrared videos. Image and Vision Computing, 29(9), 607-619.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. Cognition and Emotion, 24(8), 1377-1388.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces (KDEF). CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutes, 91(8), 630-635.

Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR), Las Vegas, NV, USA (pp. 5562-5570).

Benitez-Quiroz, C. F., Srinivasan, R., Feng, Q., Wang, Y., & Martinez, A. M. (2017). Emotionet challenge: Recognition of facial expressions of emotion in the wild. arXiv preprint arXiv:1703.01210.

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. Proceedings of the National Academy of Sciences, 111(15), E1454-E1462.

Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2584-2593).

Li, S., & Deng, W. (2018). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial

expression recognition. IEEE Transactions on Image Processing, 27(1), 348-359.

Zhang, Z., Luo, P., Chen, C. L., & Tang, X. (2018). From facial expression recognition to interpersonal relation prediction. International Journal of Computer Vision, 126(5), 1-20.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Cristinacce, D., & Cootes, T. (2006). Feature detection and tracking with constrained local models. In Proceedings of the British Machine Vision Conference (BMVC).

Saragih, J. M., & Göktaş, F. (2007). A self-tuning probabilistic model for facial feature detection and tracking in video. In Proceedings of the International Conference on Computer Vision (ICCV).

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Cao, X., Wei, Y., Wen, F., & Sun, J. (2014). Face alignment by explicit shape regression. International Journal of Computer Vision, 107(2), 177-190.

Liu, T., & Shen, C. (2019). Cascade face alignment by deep convolutional neural networks. Pattern Recognition, 95, 79-88.

Liang, Y., & Zhu, S. (2016). Local binary features for face detection and recognition. Neurocomputing, 193, 106-117.

Xiong, X., & De la Torre, F. (2013). Supervised descent method and its application to face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Dong, W., Zhu, L., Shen, J., & Shi, G. (2019). Cascaded deep neural networks for face detection and alignment. Pattern Recognition, 91, 166-174.

Ren, S., Cao, X., Wei, Y., & Sun, J. (2014). Face alignment at 3000 FPS via regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Kahou, S. E., Michalski, V., Konda, K., Memisevic, R., & Pal, C. (2016). Recurrent neural networks for emotion recognition in video. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (pp. 383- 389).

Knyazev, B., Komkov, S., Benevenuto, F., & Stokowiec, W. (2017). Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. IEEE Transactions on Affective Computing, 8(4), 626-637.

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. Entropy, 25(10), 1440.

Tong, S. G., Huang, Y. Y., & Tong, Z. M. (2019). A robust face recognition method combining LBP with multi-mirror symmetry for images with various face interferences. International Journal of Automation and Computing, 16(5), 671- 682.

Reed, S., Akata, Z., Lee, H., & Schiele, B. (2017). Learning deep representations of fine-grained visual descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(5), 1089-1101.58. arXiv:1604.04693

Aly, M., Ghallab, A., & Fathi, I. S. (2023). Enhancing Facial Expression Recognition System in Online Learning Context Using Efficient Deep Learning Model. IEEE Access.

Li, Q., Liu, X., Gong, X., & Jing, S. (2019). INDReview on Facial Expression Analysis and its Application in Education. Chinese Automation Congress.

Fang, B., Li, X., Han, G., & He, J. (2023). Facial Expression Recognition in Educational Research From the Perspective of Machine Learning: A Systematic Review. IEEE Access.

Hou, C., Ai, J., Lin, Y., Guan, C., & Zhu, W. (2022). Evaluation of Online Teaching Quality Based on Facial Expression Recognition. Future Internet.

Savchenko, A., Savchenko, L. V., & Makarov, I. (2022). Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network. IEEE Transactions on Affective Computing.

Kolkur, S., Vaghulade, S., Tejwani, G., & Vazirani, Y. (2019). Effective Classroom Monitoring by Facial Expression Recognition and Ensemble Learning. IEEE Transactions.

Zhang, L., & Tjondronegoro, D. (2011). Facial Expression Recognition Using Facial Movement Features. IEEE Transactions on Affective Computing.

Deshmukh, S. P., Patwardhan, M. S., & Mahajan, A. (2016). Survey on Real-Time Facial Expression Recognition Techniques. IET Biom..

Chopra, K., & Chitranshi, J. (2024). Effectiveness of Hybrid Learning Tools: Analysis of Engineering Colleges in India. Journal of Engineering Education Transformations, 37(4), 22–28. Scopus. https://doi.org/10.16920/jeet/2024/v37i4/24155

Lokesh, C., Shankar, S., Shilpa, R., & Rekha, K. R. (2022). Digital Transformation and Hybrid Model in Engineering Education. Journal of Engineering Education Transformations, 36(special issue 2), 93–98.Scopus.https://doi.org/10.16920/jeet/2023/v36is2/23013

Bakhare, R. (2022). Hybrid Model of Teaching-Learning is Total Chaos: An effect of Reverse Halo. Journal of Engineering Education Transformations, 36(Special Issue1), 169–184. Scopus. https://doi.org/10.16920/jeet/2022/v36is1/22189

Attieh, M., & Awad, M. (2023). Forecasting of University Students' Performance Using A Hybrid Model of Neural Networks and Fuzzy Logic. Journal of Engineering Education Transformations, 37(1), 142–156.Scopus.https://doi.org/10.16920/jeet/2023/v37i1/23140

He, J., Wen, X., & Zhou, J. (2023). Advances and Application of Facial Expression and Learning Emotion Recognition in Classroom. Proceedings of the 2023 International Conference.