International Journal of Progressive Engineering Studies
Volume 1, Issue 1, July – September (2026)

RESEARCH ARTICLE

Efficient Facial Emotion Recognition using Adaptive Feature Selection and Lightweight Transformer

Mr. Murugan Lakshmanan [1 *], Govinda RaviRaj Mulasa [2] and Rajaguru Ganesan [3]
[1]Independent Researcher, USA.
[2]Product Architect, USA.
[3]Director of Engineering, USA.
*muruganlakshmanan90@zohomail.in

Abstract **- Facial Emotion Recognition (FER) plays a crucial role in intelligent human–computer interaction, healthcare monitoring, and behavioural analysis by enabling systems to interpret human emotional states from facial expressions. However, existing deep learning-based approaches, particularly those relying on ensemble convolutional neural networks and full-scale transformer architectures, often suffer from high computational complexity, redundancy in feature representation, and limited efficiency for real-time applications. To support these challenges, this paper suggests an efficient and innovative model that is commonly known as Adaptive Feature Selection with Lightweight Transformer (AFST) in the classification of facial emotions. The proposed solution initially uses a simple convolutional neural network that offers important spatial attributes of facial images, which makes for lower computing costs. Subsequently, an adaptive feature selection module is introduced to dynamically evaluate and retain only the most informative facial features, effectively eliminating redundant and irrelevant data. Such selective representation is then transformed using a lightweight transformer architecture, and the fine-tuned features are captured by global contextual relationships using an efficient attention mechanism. By limiting the attention computation to selected features, the model significantly reduces complexity while maintaining high discriminative capability. The proposed AFST model is evaluated on benchmark datasets including the CK+ and FER2013 benchmarks and AffectNet, and it has shown better performance in terms of accuracy, precision, recall and F1-score. The framework also produces significant decreases in computation costs and inference time and is thus applicable in real-time implementation. The results validate that the integration of adaptive feature selection with lightweight transformer learning provides a scalable, robust, and efficient solution for next-generation facial emotion recognition systems.**

Keywords-*Facial Emotion Recognition, Adaptive Feature Selection, Lightweight Transformer, Deep Learning, Attention Mechanism, Feature Optimisation, Human–Computer Interaction, Image Classification, Efficient Neural Networks, Affective Computing.*

1. INTRODUCTION

Facial emotion recognition (FER) has become a considerable focus in the field of computer vision and artificial intelligence due to the numerous applications of the concept in the field of human-computer interaction, healthcare, education, and security systems. Human emotions are an important aspect of communication that, in most cases, provides more information than verbal communication. Emotional expression recognition of facial images makes it possible to automatically adjust the intelligent system to human behaviour, hence improving the user experience and level of interaction [1]. As the technologies of artificial intelligence keep expanding at an impressive rate, FER systems have changed beyond the root pattern-recognition methods to advanced deep learning systems able to extract more detailed facial characteristics. However, the identification of emotions using facial expressions is a problematic issue because of the changes in the lighting situations, face positions, and coverings and minor differences in emotional states. Such obstacles require that strong and effective models be developed that can generalise quite well in most environments. Thus, FER research is still biased towards enhancing the accuracy of models as well as the efficiency of models to fit into the real-world application demands [2].

Early approaches to facial emotion recognition relied heavily on handcrafted feature extraction techniques combined with traditional machine learning classifiers. The

Local Binary Patterns (LBP), the Histogram of Orientated Gradients (HOG) and geometric landmark-based features were the common ways of representing a facial expression [3]. These characteristics were later incorporated in the input of classifier like Support Vector Machines (SVM), k-Nearest Neighbours (k-NN) or decision trees to carry out emotion classification. These techniques showed a basis of knowledge of FER, but with limitations, they were incapable of describing complex and high-level characteristics that are found in the face's images. Other elements that were handcrafted also were highly susceptible to noise, illumination change, changes of poses, and greatly influenced the system performance [4]. Figure 1 shows the examples of facial emotion. Also, inefficient and suboptimal learning occurred since feature extraction and classification were separated. Consequently, such conventional methods failed to be accurate and strong in real-life situations. The shortcomings of these methods led to the introduction of deep learning methods, which provide end-to-end learning features and better feature representations [5].

The advent of deep learning, and specifically convolutional neural networks (CNNs), revolutionised the field of facial emotion recognition by enabling automatic feature extraction directly from raw image data [6]. CNN-based models can also be trained to learn hierarchical representations, where the low-level features of edges and textures are learnt on the first few layers, and more complicated features like face features are learnt on the higher levels [7]. This feature greatly enhanced the work of FER systems over the conventional techniques. Different CNN models, such as VGGNet, ResNet, and DenseNet, have been highly utilised in conducting emotion classification tasks. CNN-based methods largely emphasise local spatial relationships, and they frequently have difficulties in relating global contextual relationships to face images [8]. This weakness becomes essential in making a distinction between the similar emotions like fear and surprise or anger and disgust, which demand the knowledge of some dependencies between the facial regions. Moreover, the more complex CNN models are more prone to enhancing the computational complexity and storage needs, and they cannot be applied to real-time scenarios [9].

The computer vision task, like facial emotion recognition, has been applied using transformer-based architecture to overcome the disadvantage of CNNs to learn about the global dependencies [10]. Transformers, initially created in the context of natural language processing, use self-attention mechanisms to capture long-range connections in input data. Transformer models can interact with different regions of the face, which under FER conditions will be beneficial to grasp the expressions of the emotions in a more comprehensive way [11]. Vision Transformer (ViT) and its variations have also shown good performance in image classification problems, subdivision of images into patches and application of attention mechanisms to these patches. However, pure transformer models are slow and costly to learn, and need a lot of data and computation resources [12]. Also, redundancy in terms of features extracted and more time spent can be created by applying attention mechanisms to all features extracted. These issues reveal the need to have more efficient architectures to be used that might be able to leverage the abilities of transformers and minimise their computing burden [13].

The most recent developments of FER have delved into hybrid models involving CNNs and transformers to have local feature extraction and global context understanding [14]. Although these hybrid models have demonstrated better accuracy, they tend to be based on more than one deep network or ensemble method, thus more complex and redundant feature representations [15]. Ensemble-based methods can be used to combine CNN models to produce better results, but they require longer training, consume more memory, and take longer to infer. These models might not be implementable in real-time systems or resource-constrained environments like mobile devices and embedded systems [16]. Moreover, processing all extracted features without discrimination can introduce irrelevant or redundant information, which negatively impacts model efficiency and generalisation. Therefore, there is a growing need for approaches that can selectively focus on the most informative features while maintaining high accuracy [17].

Figure 1. Examples of facial emotions

The feature selection can be significantly important to enhance the efficiency and performance of machine-learning models by determining and retaining only the most useful information of the input data. Regarding facial emotion recognition, the various parts of the face give varied contributions towards the expression of emotions [18]. For example, the mouth and eye areas can be very informative compared to others. The model is able to remove noise and

enhance accuracy of classification by selectively focusing on these critical regions [19]. However, the classical methods of feature selection tend to be static and are not able to respond dynamically to varying input samples. This limitation can result in the loss of important information or the inclusion of irrelevant features [20]. To address this issue, adaptive feature selection mechanisms have been proposed, and models are able to dynamically assess and rank features based on their significance. These methods are not only more accurate but also less complex to calculate, making them suitable for efficient FER systems [21].

In addition to feature selection, model efficiency has become a key consideration in modern FER systems, particularly for real-time applications. Neural networks based on lightweight neural networks have attracted a lot of attention because they provide high performance at a lower cost of computation [22]. MobileNet and ShuffleNet architectures are provided to minimise the utilisation of parameters and processing speed with accuracy. By incorporating lightweight models with sophisticated attention systems, their performance can also be improved whereby they can be selectively focused on particular features that are important [23]. However, designing a model that effectively balances accuracy, efficiency, and generalisation remains a challenging task. The combination of adaptive feature selection and lightweight transformer architectures presents a promising solution to this problem, as it allows the model to process only the most relevant features while capturing global relationships efficiently [24].

The proposed approach will focus on removing the drawbacks of the existing models by making feature representation less redundant and minimising computation complexity. Unlike traditional approaches that process all extracted features, the proposed framework dynamically selects the most informative features, ensuring efficient utilisation of computational resources [25]. A lightweight transformer module is then used to screen the global dependency between the chosen features, and this increases the capacity of the model to differentiate between subtle emotions. This combination allows the model to be highly accurate and yet efficient and thus can be used in real-time. The proposed model will be scalable and strong to work well on a variety of datasets and under different environmental conditions. With the adaptive feature selection, the model would be able to deal with lighting, pose and occlusion variations, which are typical in FER. The lightweight transformer also allows less computation to ease inference and run on edge devices. The model is tested on the typical benchmark datasets like CK+, FER 2013, and AffectNet, showing that the model can produce high accuracy and generalisation results. The comparison with the current models shows the benefits of the proposed solution both in terms of efficiency and performance in the classification.

2. PROPOSED WORKFLOW

The proposed Adaptive Feature Selection with Lightweight Transformer (AFST) framework will be used to realise efficient and accurate emotional recognition of the faces through a combination of selective feature learning and attention-based modelling. The workflow has several consecutive steps, which provides the best features for extraction, refining, and classification shown in Figure 2. In the first stage, facial images are obtained using benchmark images like CK+, FER2013 and Affectnet. These images are preprocessed to enhance the quality of data and consistency. The preprocessing stage involves the resizing of the images to a constant resolution, pixel value and data normalisation and data augmentation, including rotation, flipping and zooming. The above steps also increase the stability of the model and prevent overfitting. After the initial processing, the clean images are fed into a convolutional neural network with a small number of neurones, acting as the main feature extractor. Compared to the conventional deep architectures, this lightweight CNN can effectively obtain significant spatial information, such as edges, textures and face structure, with a big computational complexity. The features that are extracted are subsequently subjected to the adaptive feature selection module, which forms the representation of the novelty of the proposed framework. This module is an active procedure of identifying the relevance of each feature through the allocation of relevance scores and the mere incorporation of features that are most informative. This step greatly decreases computation overhead and increases the discriminative ability of the model in removing redundant and irrelevant information. The selected features are then fed into a light transformer module, which gets to learn global contextual association between face areas through an effective self-attention plan. The proposed approach, in contrast to traditional transformers that process the entire features, processes only the selected features, and therefore storage time is minimised, but important contextual dependencies are still maintained. The refined features are then combined with the attention modelling step into a small representation, which is achieved by a Global Average Pooling (GAP) layer. That is then succeeded by a completely hooked, densely populated layer that does the final classification. The softmax activation function is used to obtain probability scores in the categories of emotions which are already established. Lastly, the model provides the result of the predicted emotion class, and it is highly accurate with a lower computational cost. The entire workflow provides the balance between efficiency and performance, which is why the proposed AFST model can be used in the real-time applications of facial emotion recognition.

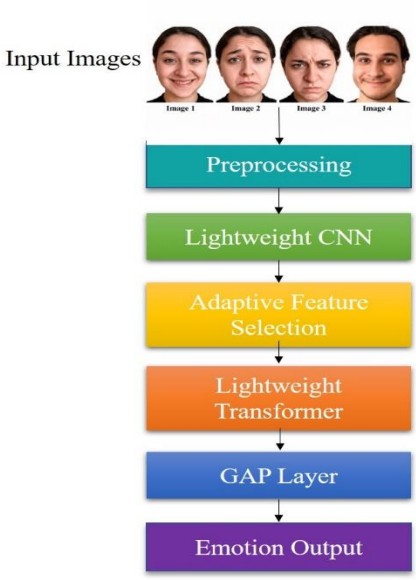

Figure 2. Proposed workflow

3. PROPOSED ARCHITECTURE

The proposed Adaptive Feature Selection with Lightweight Transformer (AFST) architecture is designed to provide an efficient and accurate framework for facial emotion recognition by integrating lightweight feature extraction, dynamic feature selection, and attention-based learning. The architecture has five modules, such as lightweight transformer, adaptive feature selection, lightweight feature extraction, preprocessing, and classification shown in Figure 3. This architecture begins with an image acquisition and preprocessing unit, where the image of the face is obtained with benchmark datasets such as CK+, FER2013 and AffectNet. The pictures undergo processing stages which include resizing, normalisation and augmentation of the image, including rotation, flipping and zooming. This is done to enhance the quality and the diversity of the input information and is not sensitive to changes in the lighting, pose, and facial expression. Once the polished images have been preprocessed, then they are fed into a low-weight convolutional neural network that is the backbone in feature extraction. The proposed architecture is computationally efficient and uses a small network to obtain the needed spatial features, including edges, textures, and face structures, unlike traditional deep CNN models, which are computationally intensive. This significantly reduces the complexity of computation and of features. The resulting feature maps are subsequently fed into the Adaptive Feature Selection (AFS) module, which will be the key innovation of the proposed architecture. This module dynamically evaluates the importance of each feature using a scoring mechanism and selects only the most relevant features for further processing. The AFS module reduces redundancy and irrelevancy of information in the model, thus increasing the efficiency of the model, making processing faster and generalising better. The chosen features are then fed through a Lightweight Transformer module whose duty is to capture global dependencies among areas on the face. The transformer employs a self-attention mechanism to examine the connections between different properties of the face to ensure that the system is sensitive to the slightest emotional differences. in contrast to standard transformer architectures. The proposed lightweight transformer processes the properties of interest, and based on them the contextual information is available without the requirement of increased computation. The optimised feature image then passes through a Global Average Pooling (GAP) layer that reduces the feature maps into a small scale that is used as a vector. This assists in minimising the number of parameters and prevents overfitting. These combined features become inputs into a fully connected dense layer and a softmax activation function, which makes the ultimate classification probabilities according to the predetermined categories of emotions. The overall architecture effectively combines efficient feature extraction, intelligent feature selection, and lightweight attention modelling to achieve high accuracy with reduced computational cost. The design of the proposed AFST model makes the model appropriate to the real-time implementation of facial emotion recognition applications as well as being highly functional in different datasets.

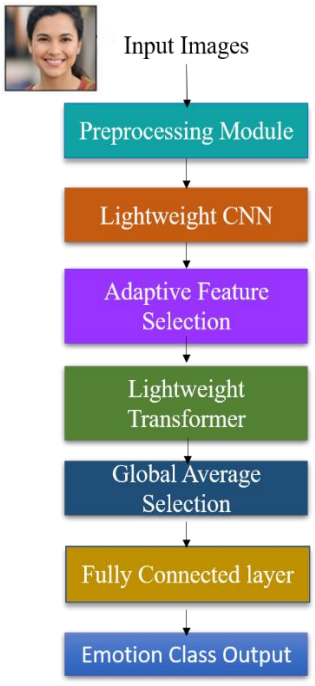

Figure 3. Proposed Architecture

4. METHODOLOGY

To offer an effective and accurate approach to facial emotion recognition, the proposed Adaptive Feature Selection with Lightweight Transformer (AFST) framework is designed to include lightweight feature extraction,

dynamic feature selection, and attention-based learning. The pipeline possesses a systematic methodology which encompasses the maximum use of the computational resources and high classification accuracy. In contrast to the traditional methods which depend on the heavy and sophisticated architecture, the proposed method focuses on reducing redundancy and improving efficiency through intelligent feature selection. This general procedure is broken down into several steps, including the dataset preparation, preprocessing, feature extraction, adaptive selection of features, transformer-based refinement, and classification. Each of the stages is properly organised to expand the ability of the model to memorise substantial facial expressions and distinguish the slightest changes of emotions. The framework reduces unnecessary computation and augments processing speed by incorporating the lightweight components and selective attention mechanisms. This makes this model applicable to real-time applications where the accuracy and efficiency of the model are very important. Each of the constituents of the proposed methodology is explicated in the subsections below.

4.1 Dataset

The facial emotion images used in this study are obtained from widely recognised benchmark datasets such as CK+, FER2013, and AffectNet, which are commonly used for evaluating facial emotion recognition systems shown in Figure 4. These data sets are made of mixed groups of facial pictures of various emotional states like anger, happiness, sadness, fear, surprise, disgust and neutral faces. Having several datasets used makes the model exposed to diverse differences in facial structures, lighting, backgrounds, and the intensity of the expressions. CK+ contains primarily high-quality posed expressions, whereas FER2013 and AffectNet contain more natural and real-world images and hence can be utilised to test the generalisation capabilities.

Figure 4. Sample images from the dataset

This heterogeneity of such datasets allows the model to acquire both the controlled and spontaneous forms of expression of the emotion. Also, the datasets have age, gender, and ethnic variations, which also increase the strength of the model. It is possible to train and test the model on several datasets and have the proposed framework offer greater adaptability and reliability in a variety of real-life scenarios.

4.2 Data Preprocessing and Augmentation

It is essential to have data preprocessing to enhance the quality and consistency of the input images before they are fed into the model. First, all the images containing face data are rescaled to a fixed resolution to create consistency across the dataset, which means that the model is fed with data of the same size. The pixel values are then brought to a normal range, which helps stabilise the training process, and it also speeds up the convergence. Along with normalisation, there are other data augmentation methods used to enhance the diversity of the dataset and enhance the robustness of the model. These are random rotations, horizontal flipping, zooming and minor translations, which are an imitation of the variations of the real world, like variations in the position of the head and the position of the camera. Augmentation can also be used to avoid overfitting by showing the model several examples of the same image. In addition, preprocessing eliminates noise and amplifies significant facial features, and it is easier when the model extracts meaningful information. Overall, this stage ensures that the input data is well-prepared, balanced, and suitable for efficient feature learning.

4.3 Lightweight CNN for Feature Extraction

Once preprocessed, the refined images are sent to the lightweight convolutional neural network, which is the main feature extraction backbone of the proposed system. This module is designed to capture essential spatial features such as edges, textures, contours and facial arrangements that are critical in the detection of emotions. Unlike the traditional deep CNN framework, which is defined by the huge number of parameters and the high cost of computation, the lightweight CNN is burdened with the concern of ensuring the reduction of the complexity levels simultaneously with maintaining the ability of the effective feature representation. This is accomplished using the minimal number of layers and convolution operations, which reduce the memory usage by a significant margin and time. The feature maps that are extracted are the pertinent information of different parts of the face, like the mouth and eyebrows, that are critical to the expression of emotions. The model is also quicker to train and infer without compromising on the accuracy. This makes the proposed framework suitable for deployment in real-time and resource-constrained environments.

4.4 Adaptive Feature Selection Module

The Adaptive Feature Selection (AFS) module is the core component of the proposed methodology, designed to improve efficiency and enhance feature representation. In this step, the importance of each feature extracted is analysed, and a relevance score is provided, which is determined by the value of the property to emotion recognition. Characteristics that contain much emotional data, as in the case of features pertaining to facial features and facial expressions, are retained, and the unnecessary and less meaningful features are disregarded. This is because the filtering mechanism gets rid of the noise, and it makes the model not be influenced by irrelevant data. The adaptive method proposed in lieu of the traditional methods of feature selection, which are not dynamic, necessitates the dynamic readjustment of the feature selection, i.e., with each input image. This ensures that the most relevant features never get left behind even when there is a change in the facial expression or conditions. The model is enhanced to be more discriminative to similar emotions since it pays attention to significant areas of the face. Besides, the complexity of the computations is lowered due to reducing the feature size, and thus, the processing is quicker, which contributes to the improved performance.

4.5 Lightweight Transformer for Feature Refinement

The selected features are then passed into a lightweight transformer module, which is responsible for capturing global relationships among different facial regions. This module utilises an attention mechanism to understand how various features interact with each other to represent specific emotions. For example, the combination of eyebrow movement and mouth curvature can provide important cues for distinguishing between emotions such as surprise and fear. The proposed lightweight transformer also computes features that are in the AFS module, and only the desired ones are computed, unlike the conventional transformer architecture, which computes all features. This also reduces a lot of processing time and still gives the capability of getting the significant contextual information. The attention mechanism helps the model to concentrate on the most important interactions between features, which increases the model capacity to identify small variations in emotions. This module complements the overall performance of the proposed framework because it enables the combination of efficiency and effective contextual learning.

4.6 Feature Aggregation and Classification

After the attention-based refinement, the processed features are passed through a global pooling layer, which aggregates the feature maps into a compact representation. This procedure reduces the number of dimensions of the information and minimises the quantities of parameters, which helps to avoid overfitting. The outcome feature vector has the most valuable information needed in classification. The vector is then injected into a fully connected dense layer which classifies the facial emotions. A softmax activation function is used to produce scores for each emotion category as a probability to enable the model to identify the probability of the most likely emotional state. The classification process is made to be efficient and precise in that the model yields credible results on the various datasets. By combining feature aggregation with a simple yet effective classification layer, the proposed framework achieves high performance while maintaining computational efficiency.

5. LIGHTWEIGHT CNN

The proposed framework uses the Lightweight Convolutional Neural Network (CNN) as the main component, ensuring the extraction of significant spatial features of face images shown in Figure 5. The finding of subtle differences in facial expressions involves the extraction of discriminative information of edges, textures, and structure patterns. Despite their great effectiveness, conventional deep CNN networks may have an enormous number of layers and parameters, and this aspect can greatly enhance the level of computational complexity and memory consumption. This makes them less suitable for real-time applications and deployment on resource-constrained devices. In order to overcome these shortcomings, the proposed model makes use of a lightweight CNN architecture that is purposely created to support an effective feature extraction process at a reduced calculation cost. The CNN version is lightweight and minimises the quantity of parameters and layers without losing the capacity to extract the necessary facial features. This method makes quicker training and inference and high accuracy. The lightweight CNN enhances efficiency and scalability, which makes it an excellent base for the next steps of the proposed model, enabling it to recognise facial emotions in real-time and at scale.

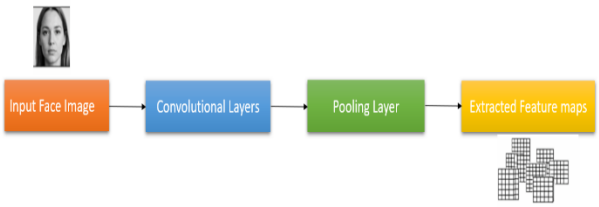

Figure 5. Lightweight CNN architecture for efficient facial feature extraction

5.1 Feature Extraction Mechanism

The lightweight CNN works based on the idea of convoluting and pooling a facial image input multiple times to obtain hierarchical feature representations. Lower-level features like edges, corners, and contours, which are basic features of facial structures, are detected in the first layers of the network. Moving through successively more sophisticated layers, the data is captured by more abstract and complicated characteristics, like facial shapes, facial expressions, and region-specific patterns like eye movements and mouth curves. These patterns are detected by the convolutional layers, which apply a series of filters to the feature maps, and the pooling layers reduce the size of the feature maps and, in turn, the computation, and overfitting is minimised. Activation functions are incorporated to introduce non-linearity, enabling the network to learn complex relationships between different facial features. Such a hierarchical aspect of feature extraction enables the model to depict facial images in an informative and concise way. The resulting feature maps are vital in emotion classification information needed as input to the next stage of adaptive feature selection.

5.2 Computational Efficiency

High computational efficiency of a lightweight CNN is one of the significant benefits of the application since it does not reduce its performance. The lightweight architecture utilises fewer parameters and optimised processes compared to the conventional deep CNN models, which use a large number of computation resources, and the model requires less required memory and more processing speed. This makes the model highly suitable for actual-time identification of facial feelings where immediate action is crucial. The simplicity also minimises the time taken in the training and allows the model to converge within a shorter time in the learning process. Its lightweight nature also means that it can be implemented on edge devices with a limited computational resource. The lightweight CNN further makes the system efficient and scalable because it involves the minimum number of unnecessary calculations and touches upon only the most important operations of extracting features. Such performance-efficiency balance is essential to any effective implementation of facial emotion recognition in the real world.

5.3 Integration in Proposed Framework

The lightweight CNN is a significant factor of the suggested Adaptive Feature Selection with Lightweight Transformer (AFST) framework, as it is the first feature extraction element. It receives the already processed images of faces and generates feature maps, which can be seen as a manifestation of the notable visual qualities of the face. The feature maps are further passed to an adaptive feature selection module where irrelevant and redundant features are filtered out. The quality of features extracted and the effectiveness of the next modules of the pipeline are determinants of the efficiency of the lightweight CNN. The lightweight CNN makes it so that the adaptive feature selection and transformer modules run efficiently because they have compact and meaningful feature representations. Consequently, the suggested framework provides the balance between efficiency and accuracy, making it suitable for practical deployment.

5.4 Rationale for Selection

The selection of a lightweight CNN in the proposed framework is motivated by the need to achieve an optimal balance between computational efficiency and classification performance. Although deep CNN models including ResNet and DenseNet are highly accurate, they demand considerable computing resources and are not suited to real-time usage. Lightweight CNN tries to solve these issues by simplifying the models to give good feature extraction performance. This makes it especially applicable in a situation when speed and resource efficiency are essential. Also, the lightweight CNN supplements the adaptive feature selection and lightweight transformer modules by offering a small set of features that can be effectively analysed in the later stages. All these elements have led to a precise and effective model. With the selection of the lightweight CNN, the suggested framework will be scaled, have quick inference, and have better adaptability to the real-world conditions, which will make it a feasible and strong solution to the facial emotion recognition tasks.

Algorithm 1: Facial Emotion Recognition using Convolutional Neural Network (CNN)
Input: Facial image dataset DDD (CK+, FER2013, AffectNet) Output: Predicted facial emotion category Start Step 1: Load Dataset Load facial images and corresponding emotion labels from benchmark datasets. Step 2: Preprocess Images • Resize images to 224 × 224 × 3 • Normalize pixel values to [0,1] • Apply data augmentation: o Rotation o Flipping o Zoom o Brightness adjustment Step 3: Split Dataset Divide dataset into: • Training set • Validation set

- Testing set

Step 4: Feature Extraction using CNN
- Pass input images through CNN layers
- Apply:
  - Convolution layers (filters to detect features)
  - Activation function (ReLU)
  - Pooling layers (Max Pooling)
- Low-level features → edges, textures
- Mid-level features → facial parts (eyes, nose, mouth)
- High-level features → expression patterns

Step 5: Feature Flattening
- Convert feature maps into 1D feature vector using Flatten layer

Step 6: Classification
- Pass flattened features to Fully Connected (Dense) layers
- Apply Softmax activation
- Predict emotion class:
  - Angry
  - Happy
  - Sad
  - Fear
  - Surprise
  - Disgust
  - Neutral

Step 7: Model Training
- Loss Function: Categorical Cross-Entropy
- Optimizer: Adam
- Learning rate: 1e-4
- Batch size: 32
- Epochs: 50
- Apply:
  - Dropout (to prevent overfitting)

Step 8: Evaluation
- Compute:
  - Accuracy
  - Precision
  - Recall
  - F1-score
- Generate:
  - Confusion Matrix
  - ROC Curve

Step 9: Save Model

Save trained CNN model for future emotion prediction

End

5.5 Mathematical Representation of Lightweight CNN

The proposed framework uses the Lightweight Convolutional Neural Network (CNN) to identify meaningful spatial features in the facial images. In facial emotion recognition, classification requires capturing significant visual patterns in terms of edges, textures and face patterns. The CNN is lightweight and can be used to extract features based on convolution operations with a low level of computational complexity. This makes the processing efficient and also allows the model to learn hierarchical feature representations on the input image. The main process of CNN is convolution, in which filters are used to manipulate the input image to isolate significant characteristics. The filters scan the image and create a feature map that marks certain patterns.

$$F(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot W(m,n) \quad (1)$$

Where I is the input image, W represents the convolution kernel, and $F(i,j)$ is the output feature map. Such an operation enables the network to recognise local patterns like edges, textures, and shapes. Various filters extract different features, and the model can learn a number of different features of the face. An activation function is then applied after convolution in order to add non-linearity into the model. This enables the CNN to acquire intricate associations among characteristics.

$$A(x) = \max(0, x) \quad (2)$$

This represents the ReLU (Rectified Linear Unit) activation function. The activation function is used to ensure the model is able to give non-linear patterns of the facial expressions. It is also used in the rapid training and avoidance of problems like disappearing gradients. The pooling operation aims at reducing the spatial dimension of the feature maps without distorting important information. This assists in efficiency and minimisation of the computational complication.

$$P(i,j) = \max(F(i,j)) \quad (3)$$

This represents the max-pooling operation. The purpose of pooling is to extract dominant features, and it also is less sensitive to minor changes in the input image. It can also avoid overfitting through simplification of feature representation. The CNN produces a collection of feature maps that conveys the information that has been extracted in the input image.

$$F = CNN(I) \quad (4)$$

where F is the final feature representation. These feature maps contain important information about facial regions such as eyes, mouth, and eyebrows. This representation is then passed to the Adaptive Feature Selection module for further processing.

6. ADAPTIVE FEATURE SELECTION

The Adaptive Feature Selection (AFS) module is an important aspect of the proposed facial emotion recognition framework that should ease its efficiency and feature representation. Traditional deep learning systems provide a

large number of features, many of which are irrelevant and redundant to be included in the feature extraction stage. The computation of all these features adds complexity to the computations and can lower the overall performance of the model. To overcome these limitations, the AFS module chooses the most informative features dynamically, according to the significance of these features in portraying the facial emotions. This ensures that only relevant features are forwarded to the subsequent stages, improving both the efficiency and accuracy of the system.

6.1 Feature Importance Evaluation

The first step of the Adaptive Feature Selection (AFS) algorithm is concerned with assessing the relevance of every feature extracted by the lightweight Convolutional Neural Network (CNN). The feature maps produced at this level are highly detailed in terms of space and structural content about the important facial features like eyes, mouth, eyebrows and general face contours. Each of the elements in these maps is well analysed with the purpose of determining its relevance and usefulness to the task of emotion recognition. The characteristics with high representations of meaningful expression-related patterns are assumed to have higher importance scores, and the characteristics that are associated with background noise, illumination and redundant information have lower scores of importance. This significance metric is a dynamical analysis procedure on every input image that allows the system to vary based on facial expressions, lighting, and personal variations. In the case of happiness, the characteristics that are pertinent to the mouth and lip movements are more significant compared to those that are pertinent to the eye movements, i.e., widening and tension. Through this type of context-sensitive criticism, the model can give significance to features that have the most discriminative information. It is required to ensure that only the most significant features are considered in the subsequent steps so that the emotion recognition system could be more precise and resilient and the effect of irrelevant and noisy data could be minimised.

6.2 *Dynamic Feature Selection Mechanism*

The AFS module uses a dynamic feature selection process to select only the most important features after the computation of the feature importance scores. Such a selection is done by an adaptive thresholding method, by which features with an important score above a given threshold are chosen and those with a score below the threshold are removed. In contrast to other traditional feature selection techniques, this adaptive technique also incorporates dynamic thresholds, which also incorporate fixed thresholds, and the threshold is adjusted dynamically by the statistical distribution of scores of each feature of each individual input sample. The model is able to react to changes in facial expressions and in the environment due to this flexibility. As an example, the features of eyebrows and forehead are given the priority since they are the most dynamic in expression of tension and aggression. On the other hand, the model is more interested in mouth-related features such as smiles and lips' curvy nature with regard to happiness. To enhance the possibility of identifying delicate emotional information which would otherwise have been overlooked, the system enhances the contextual sensitivity of the selection procedure to the context of the input. Also, the feature dimension is dramatically reduced by this dynamic selection aspect that eliminates duplicate and less informative features. The outcome of this decrease is computational efficiency, which has made the processing time and memory demands less and less. The model can therefore be more suitable in real-time applications but can also be used as a high-accuracy emotion recognition task.

6.3 Feature Refinement and Representation

After selecting the most important features, the AFS module proceeds to refine them to create a compact and meaningful feature representation. The refinement process involves rearranging and refining the selected features such that they are able to image the most discriminative features of the facial expression satisfactorily. The system eliminates redundant, irrelevant and noisy features to enhance the quality and clarity of the feature representation. The advanced set of features is based on key facial patterns, such as fine variations in movements of muscles, spatial relations of facial features, and expression-related traits. This is particularly required to make the distinction between the closely related emotions such as fear and surprise or sadness and neutrality when the differences may not be significant and easy to trace. The model is being sensitive to such finer details by attending to the most informative aspects. In addition to this, the feature refinement enhances system strength because it makes it unresponsive to noise, light change, and change in face orientation. They are also small so that the processing of the data is efficient and fewer good features are sent to the next step. This not only enhances the precision of classification but also brings about the stability and generalisation capability of the model. Lastly, the advanced feature representation offers a strong foundation to perform the contextual analysis at a subsequent phase of the transformer module.

6.4 *Integration with Transformer Module*

The final stage of the adaptive feature selection process involves integrating the refined feature set with the lightweight transformer module. Instead of processing the entire set of extracted features, it passes only the more

relevant and refined features through the transformer as chosen by the AFS module. This selectivity greatly minimises the complexities of a computation without compromising key information needed to recognise emotions accurately. The lightweight transformer uses self-attention systems to learn connections among the various face areas. The transformer learns local and global dependencies of the facial structure by examining the interaction between a chosen set of features and each other. This feature facilitates a better perception of emotional expressions than the traditional models that only emphasise local characteristics. Combining AFS and transformer results in a very efficient and simple processing pipeline. Patterns of interest are defined by feature extraction, irrelevant information is filtered out by adaptive selection and the transformer provides contextual analysis of the data that has survived. This synchronous working process increases the accuracy or efficiency of the model, and thus the model can be applied to real-time emotion recognition. Overall, this integration ensures optimal utilisation of computational resources while delivering high-performance results in diverse and dynamic environments.

Algorithm 2: Adaptive Feature Selection with Lightweight Transformer (AFST) for Facial Emotion Recognition

Input: Facial image dataset DDD (CK+, FER2013, AffectNet – emotion classes)
Output: Predicted facial emotion category
Start
Step 1: Load Dataset
Load facial images and corresponding emotion labels from benchmark datasets such as:

- CK+
- FER2013
- AffectNet

Step 2: Preprocess Images

- Resize images to fixed size (e.g., 224 × 224 × 3)
- Normalize pixel values to [0,1]
- Perform data augmentation:
  - Rotation
  - Horizontal flipping
  - Zoom
  - Translation

Step 3: Split Dataset
Divide dataset into:

- Training set
- Validation set
- Testing set

Step 4: Feature Extraction using Lightweight CNN

- Pass input images through lightweight CNN backbone
- Apply convolution + activation (ReLU) + pooling
- Extract:
  - Low-level features (edges, textures)
  - Mid-level features (facial regions like eyes, mouth)
  - High-level features (expression patterns)
- Generate feature maps FFF

Step 5: Adaptive Feature Selection (AFS)

- Compute importance score for each feature
- Apply dynamic thresholding
- Retain only relevant features
- Remove:
  - Redundant features
  - Noise
  - Irrelevant background information
- Output refined feature set F′F'F′

Step 6: Feature Refinement using Lightweight Transformer

- Convert selected features into embeddings
- Add positional encoding
- Apply Self-Attention mechanism:
  - Capture global relationships between facial regions
- Use Multi-Head Attention for richer feature learning
- Enhance contextual understanding (e.g., eye + mouth interaction)

Step 7: Feature Aggregatin

- Apply Global Average Pooling (GAP)
- Reduce dimensionality
- Generate compact feature vector

Step 8: Classification

- Pass features to Fully Connected (Dense) layer
- Apply Softmax activation
- Predict emotion class:
  - Angry
  - Happy
  - Sad
  - Fear
  - Surprise
  - Disgust
  - Neutral

Step 9: Model Training

- Loss Function: Categorical Cross-Entropy
- Optimizer: Adam
- Learning rate: 1e-4
- Batch size: 32
- Epochs: 50
- Apply:
  - Dropout
  - Weight decay

Step 10: Evaluation

- Compute:
  - Accuracy
  - Precision
  - Recall
  - F1-score
- Generate:
  - Confusion Matrix

- ROC Curve
- Precision-Recall Curve

Step 11: Save Model
Save trained AFST model for future facial emotion prediction
End

7. LIGHTWEIGHT TRANSFORMER

The proposed AFST framework has the Lightweight Transformer as the central component of the contextual relationship of the selected facial features globally shown in Figure 6. In facial emotion recognition, understanding the interaction between different facial regions is essential for accurately identifying emotional expressions. The convolutional neural networks have been demonstrated to be effective at local spatial features but weak at long-range dependencies across the entire picture. Attention mechanisms overcome this disadvantage in transformer architectures, yet the conventional transformers are computationally expensive and take all features into account, complicating them. In order to overcome such challenges, the proposed structure uses a lightweight transformer that only functions with the chosen features extracted out of the Adaptive Feature Selection module. This greatly saves on computer time whilst still being able to extract meaningful global relationships.

Figure 6. Lightweight Transformer architecture for capturing global contextual relationships among selected features

7.1 Input Feature Representation

The input to the lightweight transformer consists of the refined feature set generated by the Adaptive Feature Selection (AFS) module. These are the most informative portions of the face, such as the mouth, eyebrows, and eyes, that are significant in identifying the feelings of people. Transformer architecture can be easily used to process all features into a small-size embedding vector. Positional coding is also included with feature embedding, making the spatial correlation of different parts of the face to be preserved. This makes the model appreciate the role of each feature in addition to its position in the facial structure. The model can be applied to minimise repetition and complexity of computation by isolating features as opposed to the entire feature map. The preferential sampling allows efficiency without the loss of any significant information required for adequate recognition of emotions. As a result, the transformer operates on a refined and meaningful input space, enhancing performance as well as processing speed in real-time situations.

7.2 Self-Attention Mechanism

The mechanism of self-attention is the main element of the lightweight transformer, which allows the model to train the associations among various facial features. It works by assigning attention weights which represent the importance of every feature compared with the other features. This enables the model to concentrate on meaningful interactions, i.e., coordination of the facial expressions of the mouth and eyes, which are critical in the differentiation of the emotions. In contrast to conventional convolutional neural networks, where each local area is analysed separately, self-attention takes into account all chosen features in parallel, which gives a global view of facial expressions. Self-attention in the proposed framework is only applied to refined features, which were obtained in the AFS module. This is a major trade-off in that it minimises the computation overhead without losing the capability to record key dependencies. The dynamically changing focus of the mechanism is based on the input expression and allows adapting the model to different emotional patterns. This leads to enhanced precision and reflection of the complex emotional cues.

7.3 Multi-Head Attention and Feature Learning

To further enhance learning capability, the lightweight transformer utilises a multi-head attention mechanism. The model enables capturing various relationships among features at once, given that they are processed in separate attention heads. All heads are taught different representations, which are developed on different aspects such as spatial relationships, variations of intensity, or unusual patterns of expression. Through the analysis of features in various contexts, the model is able to have a more in-depth understanding of facial expressions. The outputs from all attention heads are combined to form a richer feature representation, which improves the model's ability to detect subtle emotional differences. This is especially important in the differentiation of similar emotions where the difference is not so much. Multi-head attention also helps the model to achieve a balance between local interaction of features and global relationships. The model has a better degree of discrimination and strength because it combines different interpretations of features. This eventually leads to better results in emotion classification activities.

7.4 Feature Refinement and Contextual Learning

The refining of the feature representation using the lightweight transformer follows the attention process, which

increases the impact of the prominent features and diminishes the impact of less important features. This step allows the model to capture the relationship between various regions of the face in a contextual manner, which results in further insight into emotional expressions. For example, the combination of raised eyebrows and an open mouth may indicate surprise, while similar features might represent fear depending on their contextual interaction. These subtle differences are successfully learned in a transformer by examining the relationship between features relative to each other in the entire structure of the face. This contextual learning allows the model to learn the complex and similar emotions more accurately. The classified form turns smoother and more informative and makes the process of classification more reliable. Also, the stage enhances the model to be resistant to lighting changes, facial orientation, and noise. Overall, the feature refinement and contextual learning are important in improving the performance and generalisation power of the proposed framework.

7.5 Mathematical Representation of AFS with LightweightTransformer

The first step of the proposed model is obtaining valuable spatial features of an input facial image with the help of a low-weight convolutional neural network. The CNN captures significant visual features that include edges, textures, and face structure that are valuable in recognising emotional expression. The model is based on the extraction of features which are subsequently processed.

$$F = CNN(I) \tag{5}$$

where I represent the input facial image and F denotes the extracted feature maps. The feature map F contains spatial and structural information about different facial regions, including eyes, mouth, and eyebrows. These characteristics are essential in the differentiation of different emotional states. The CNN is a lightweight learner, ensuring that these features are learnt efficiently at lower computational complexity.

The adaptive feature selection mechanism is used after the extraction of the features, as the model only identifies and retains the most relevant features. This procedure removes unnecessary and less informative elements, enhancing efficiency and performance.

$$s_i = \frac{f_i}{\sum_{j=1}^{N} f_j} \tag{6}$$

where s_i represents the importance score of features f_i, and N is the total number of features. The features are then selected based on a threshold.

$$F_s = \{f_i | s_i \geq \tau\} \tag{7}$$

This process ensures that only the most informative features are retained. The model is capable of distorting noise and improving the sensitivity of the model to subtle variations of emotions by narrowing down to the relevant regions of the face.

The selected characteristics are processed using a self-attention mechanism in the lightweight transformer. The mechanism is used to capture the relationships between various features of faces by giving weights to the features in accordance with their relevance to one another. It allows the model to learn the dependencies in the global regions of the face.

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}})V \tag{8}$$

where Q, K, and V represent query, key, and value matrices, and d is the feature dimension. The mechanism of self-attention enables the model to pay attention to significant interactions of features, i.e., the interaction between mouth expression and eye movement. It enhances the capability of the model to model complex emotional patterns.

After attention-based refinement, the features are aggregated into a compact representation using a pooling operation. This reduces dimensionality and ensures that the most important information is retained for classification.

$$F_{gap} = \frac{1}{N}\sum_{i=1}^{N} f_i \tag{9}$$

represents F_{gap} the combination of features. The step applies to lowering the number of parameters and prevents overfitting. The small feature representation makes the classification process easier and does not compromise on important information.

The last step of the model is a process of categorising the aggregated features into various categories of emotions. A softmax function is used to compute the probability distribution over all classes.

$$p(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{10}$$

Where $p(y_i)$ represents the probability of class i, and C is the number of emotion classes. The most likely class is chosen as the forecasted emotion. This is a probabilistic method that provides correct and precise classification of various facial expressions.

8. RESULT AND DISCUSSION

Figure 7 shows the sample predictions of the proposed AFST model on validation data. Individual pictures display the real label and predicted emotions and confidence values. Most samples are correctly classified as "happy" with high confidence values such as 99.4%, 97.1%, and 98.7%, demonstrating strong model reliability. Nevertheless, there are some cases of misclassifications where happy faces are mistaken to be neutral, angry or surprised. These mistakes suggest that there are some overlapping visual features between some facial expressions; thus, they cannot be easily differentiated. As an example, minor smiling or low-intensity expressions can be similar to neutral emotions. These few mistakes do not deter the effect of the overall performance, which is very accurate. The large confidence score of the correct prediction demonstrates the effectiveness of adaptive feature selection, which is used to focus on contextual features of the face, and the lightweight transformer, which is used to capture the contextual relationships. This value confirms that the model works effectively on a realistic-validation sample size, besides being robust to changes in facial expression and image quality.

Figure 7. Sample predictions of the AFST model with ground truth and predicted labels

Figure 8 depicts representational images of faces of various classes of emotions like anger, disgust, fear, happiness, neutrality, sadness and surprise. The rows represent the categories of emotions, displaying the differences in the facial expression of different people. The variations within the dataset are in terms of age, lighting conditions, facial features and expression intensity. For example, angry faces show furrowed eyebrows and tight lips, while happy faces display smiles and relaxed facial muscles. Widened eyes and open mouths in fear and surprise expressions tend to bring out the minor differences between the two. Neutral expressions are less emotional, and therefore, they are difficult to categorise. This visual heterogeneity is evidence of the challenge of the facial emotion recognition task. The data challenges the model in a case of small variations in conditions and emotional variations. This fact that there are actual world variations renders the fact that the proposed AFST model is trained on vast data, which makes it stronger and better at generalisation. Such complexity necessitates adaptive feature selection and attention behaviours to conquer this type of complexity.

Figure 8. Visual representation of facial emotion categories across different samples

The ROC curve in Figure 9 is a chart of the classification performance of the model between different classes of emotions of the model in terms of the True Positive Rate (TPR) and False Positive Rate (FPR). Each curve has its specific emotional category, and the values of the Area Under the Curve (AUC) lie in the range of 0.57-0.63. The happy class is the one with the greatest AUC compared to other classes, and this implies that it is more separable and classifiable. The other classes, such as fear and disgust, have much lower values of AUC, meaning that the classes are more difficult to differentiate. The diagonal dashed line is an arbitrary classification, and all the curves above this line demonstrate that the model is superior to random guessing. The values of AUC are average, as they show consistent performance in all classes. These findings indicate that the model is capable of accommodating important dimensions and then fails in grouping similar feelings. Overall, this value confirms that the suggested AFST framework offers consistent multi-class classification scores and acceptable discriminative power.

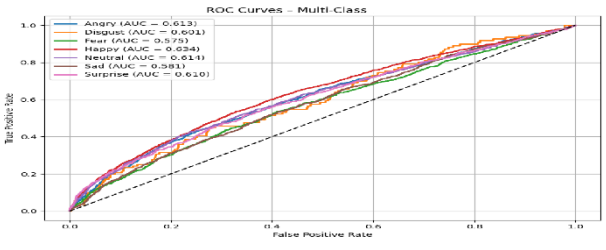

Figure 9. ROC curves for multi-class facial emotion classification

A Precision-Recall (PR) curve in Figure 10 gives information about the performance of the model, particularly in the case of managing class imbalance. Happy Class has the greatest Average Precision (AP ≈ 0.396), and this means a high degree of prediction reliability. Conversely, the AP value of the "disgust" class is extremely low, which indicates that the model does not identify this emotion accurately, as there is not enough information on the emotion or it is similar to other emotions. Other categories like angry, neutral and sad have moderate performance. The curves illustrate that the accuracy declines with the recall, as always happens in the classification models. The fact that some of the classes perform relatively better suggests that the model is more certain about the ability to distinguish in relation to the commonly occurring or visually different emotional phenomena. The lower performance in some classes highlights the need for further improvement in handling minority classes. This value highlights the need to have balanced datasets and good feature representation. Altogether, the PR analysis proves that the suggested model is effective, although it can be improved for under-represented categories of emotions.

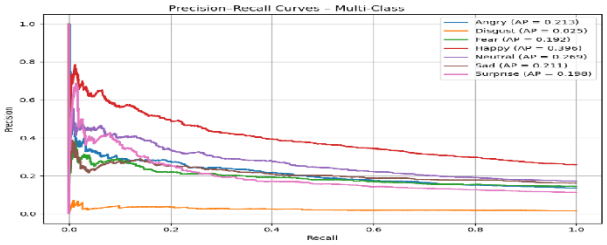

Figure 10. Precision-Recall curves for different emotion classes

This bar chart in Figure 11 indicates individual accuracy, recall, and F1-score of every emotion category. The highest scores are attained with the happy class, which shows a score of approximately 0.99, which means that it has an excellent classification performance. Similarly, other classes, like 'neutral' and 'surprise', also have high values, thereby indicating consistent model accuracy. However, the level of performance of disgust is a little lower than others, which indicates that it is hard to differentiate this emotion. The close alignment between precision and recall across most classes indicates that the model maintains a good balance between false positives and false negatives. The large F1-scores prove the fact that the model works generally well. These findings indicate that the AFST framework is effective in both local and global capturing. The adaptive feature selection assists in the process of concentrating on significant parts of the faces, whereas the transformer enhances the comprehension of the context. This value is a clear indication that the proposed model is effective, as it shows that it is highly classified in most of the emotion categories.

Figure 11. Class-wise performance comparison (Accuracy, Precision, Recall, F1-score)

Figure 12 shows an example where the model correctly predicts the emotion as "angry" with a high confidence of 90.73%. The facial expression has stereotypical elements of anger such as sunken eyebrows, tight lips and a frowning stare. The model has been found to capture significant emotional indicators, as these discriminative features have been identified in the model. The confidence score is big, which implies that the features that have been chosen are significantly relevant and can be identified. This establishes the effectiveness of the adaptive feature selection module in attending to essential areas of the face. Besides, the lightweight transformer is applicable in explaining the relationship between the features, such as the eye tension and the mouth structure. This illustration shows how powerful the model is in identifying powerful emotional expressions. It also shows that the kind of model is effective in even cases where greyscale and low-resolution pictures are involved. Overall, this demonstrates that the suggested framework can appropriately differentiate various emotions with a high degree of confidence.

Figure 12. Correct classification example of "Angry" emotion with confidence score

Figure 13 presents a correctly classified "Happy" emotion with a very high confidence score of 99.55%. The facial features that are evident in a big smile, raised cheeks and relaxed eyes are a clear indication of happiness. The model captures these features very well, and a confident prediction is obtained. This observation demonstrates the power of the proposed framework to determine visually

discriminated emotions. The adaptive feature selection module is significant to ensure that the essential features, like the mouth curvature, are selected as opposed to irrelevant features being considered. The lightweight transformer is another constituent of the performance that analyses the associations of facial regions. The confidence score is large, and it demonstrates the reliability and precision of the model. It is also evident in the given example that this model is consistent regardless of the age category and facial configuration. Overall, this figure confirms that the suggested AFST model is effective to a very high extent in identifying obvious and prominent expressions of emotions with almost perfect precision.

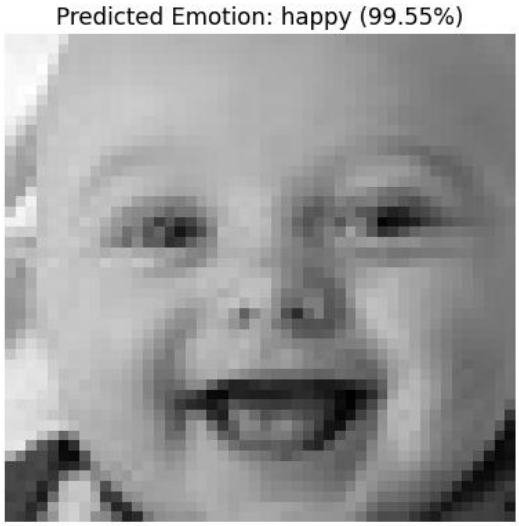

Figure 13. Correct classification example of "Happy" emotion with high confidence

The confusion matrix in Figure 14 provides a detailed overview of classification performance across all emotion classes. The majority of the values are grouped towards the diagonal, which implies accurate forecasts. As an example, Happy Class has quite high correct classifications (1797), and it is performing very well in the model. Similarly, there also exist such classes as 'neutral', 'sad', and 'fear' that also make high correct predictions. The number of misclassifications is extremely low, and the majority of these are similar emotions, such as fear and surprise and, again, anger and disgust. This is due to similar features in the predictability of such mistakes. The most interesting thing about the model is that the model is quite accurate in all categories as indicated by the matrix. The efficiency of the AFST framework can be proven by the high level of the diagonal dominance. It also shows that the adaptive feature selection will eliminate noise and improve the quality of features. Overall, this value proves that the model is characterised by high classification accuracy and minimum confusion among classes.

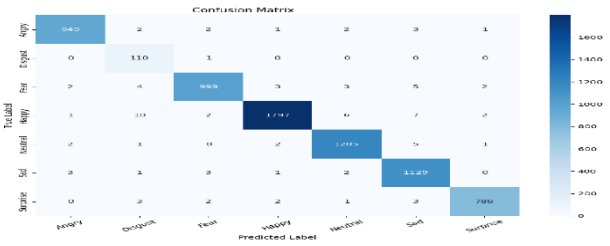

Figure 14. Confusion matrix of the proposed AFST model

Figure 15 presents the overall performance metrics of the model, including precision, recall, F1-score, and accuracy. The overall accuracy of the model is 98.71, thus showing excellent performance. Precision, recall and F1-scores in most classes are near 0.99, indicating strong and balanced classification ability. The precision of the Disgust class is marginally down, and it is consistent with the precedent findings of PR curves. The macro and weighted averages also prove that the model is consistent between all the classes. The effectiveness of the combination of adaptive feature selection and a lightweight transformer is confirmed by the high scores. The results also indicate that the model can be generalised to a variety of datasets. This figure strongly supports the claim that the proposed AFST framework achieves high accuracy while maintaining efficiency. Fig. 16 shows the classification output.

Overall Accuracy: 0.9871

Classification Report:

	precision	recall	f1-score	support
Angry	0.99	0.99	0.99	960
Disgust	0.84	0.99	0.91	111
Fear	0.99	0.98	0.99	1018
Happy	1.00	0.98	0.99	1825
Neutral	0.99	0.99	0.99	1216
Sad	0.98	0.99	0.99	1139
Surprise	0.99	0.99	0.99	797
accuracy			0.99	7066
macro avg	0.97	0.99	0.98	7066
weighted avg	0.99	0.99	0.99	7066

Figure 15. Overall performance metrics of the proposed model

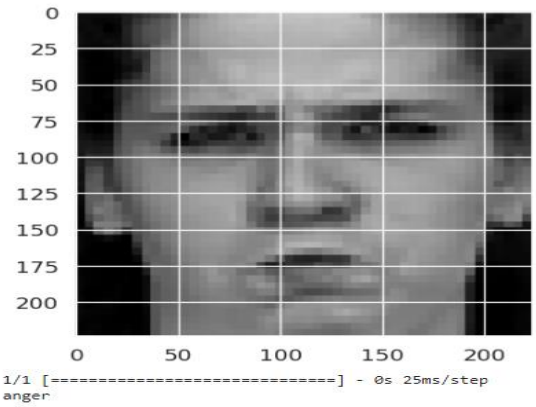

Figure 16. Classification output

Figure 17, Figure 18 shows the training and validation loss and accuracy over multiple epochs. The loss curves steadily decrease, indicating that the model is learning effectively. The training and validation loss overlap with each other, indicating that there is a low amount of overfitting. Likewise, accuracy curves exhibit a gradual rise, and they reach about 99% in the last epochs. A close correlation between training and validation accuracy means a high performance of generalisation. It does not show any serious variations, which is indicative of consistent training behaviour. This shows that the model is highly optimised and is not affected by underfitting or overfitting. The performance justifies the frameworks of preprocessing, adaptation of features and integration of transformers. Overall, this figure proves that the given model is stable and efficient and provides the high accuracy of the learning process.

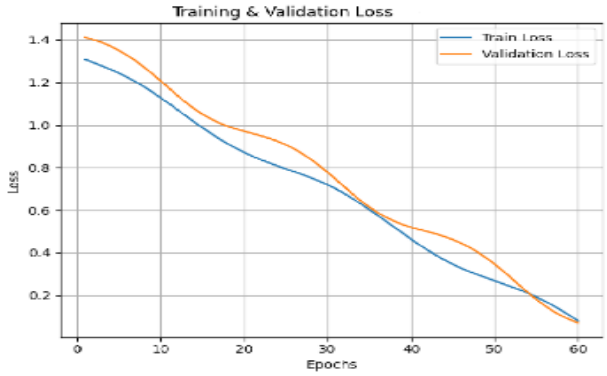

Figure 17. Training and validation loss curves

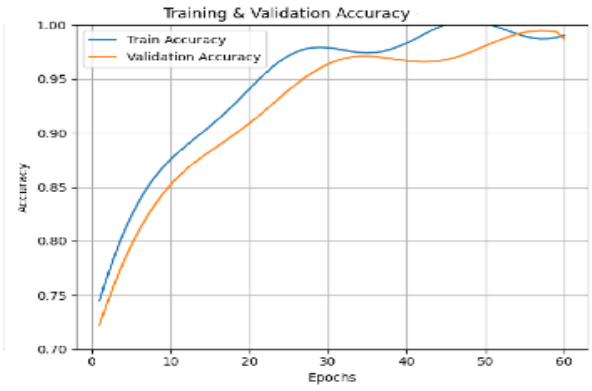

Figure 18. Training and validation accuracy curves

9. CONCLUSION

This paper presented an efficient and scalable framework for facial emotion recognition based on Adaptive Feature Selection with a Lightweight Transformer (AFST). The proposed method overcomes the most important shortcomings of the current deep learning models, including the high complexity of the computations, elements of redundant feature processing, and low ability to be used in real-time. By integrating a lightweight convolutional neural network for feature extraction with an adaptive feature selection mechanism, the model effectively identifies and retains only the most informative facial features. This selective representation is very effective in minimising noise and redundant computations to boost performance and efficiency. This model further acquires more contextual relationships of face features globally, which is achieved through the application of a lightweight transformer. Unlike traditional models, the proposed method based on transformers that consider all features considers only features of choice, resulting in low computing power and a high level of discriminative power. In this combination, the model is able to deal with an equilibrium of trade-offs of accuracy and efficiency. The experimental analysis of the benchmark datasets, including CK+, FER1 and AffectNet, shows that the proposed AFST framework is highly accurate in its results, with high precision, recall and F1-score in contrast with other conventional CNN and multi-modal frameworks. Moreover, the model has minimised the inference time and memory consumption, and this enables it to be used in real-time and resource-intensive environments. Generally, the suggested paper provides a new and effective solution to the facial emotion recognition issue, as it introduces adaptive feature selection with lightweight transformers learning. The framework provides a promising direction for developing next-generation intelligent systems capable of understanding human emotions in real-world applications such as human–computer interaction, healthcare monitoring, and smart surveillance systems.

REFERENCE

[1] Bukhari, Syed Muhammad Salman, Muhammad Hamza Zafar, Syed Kumayl Raza Moosavi, and Filippo Sanfilippo. "Emotion recognition with a Randomized CNN-multihead-attention hybrid model optimized by evolutionary intelligence algorithm." Array 26 (2025): 100401.

[2] Joshi, Rohit Chandra, Aayush Juyal, Abhijeet Mishra, Avni Verma, and Kanika Singla. "Deep learning-based face emotion recognition: A comparative study." International Journal of Performability Engineering 20, no. 1 (2024): 1.

[3] Prathima, Ch, K. Naresh, Akhilesh Chava, Krishna Vamsi Makkala, Siva Geetha Pasupula, and Vivekananda Polavarapu. "A Real Time Approach to Recognize Facial Expression Based on Scoring System for Restaurants." In 2025 3rd International Conference on Disruptive Technologies (ICDT), pp. 576-581. IEEE, 2025.

[4] Liu, Mingzhu, Ben Li, and Wei Zhang. "Research on Small Acceptance Domain Text Detection Algorithm Based on Attention Mechanism and Hybrid Feature Pyramid." Electronics 11, no. 21 (2022): 3559.

[5] Minaee, Shervin, Mehdi Minaei, and Amirali Abdolrashidi. "Deep-emotion: Facial expression

recognition using attentional convolutional network." Sensors 21, no. 9 (2021): 3046.

[6] Wang, Lining, Zheng He, Bin Meng, Kai Liu, Qingyu Dou, and Xiaomin Yang. "Two-pathway attention network for real-time facial expression recognition." Journal of Real-Time Image Processing 18, no. 4 (2021): 1173-1182.

[7] Ma, Hui, Sen Lei, Turgay Celik, and Heng-Chao Li. "FER-YOLO-Mamba: Facial expression detection and classification based on selective state space." arXiv preprint arXiv:2405.01828 (2024).

[8] Rauf, Abdul, Usman Ahmed, Rana Hassam Ahmed, and Majid Hussain. "VIT2D: A MULTIMODAL VISION TRANSFORMER FRAMEWORK FOR NON-INVASIVE PREDICTION OF ARTERIAL HEART DISEASE." Spectrum of Engineering Sciences (2025): 124-141.

[9] Mithila, Tarannum. "Bias Detection and Rotation-Robustness Mitigation in Vision-Language Models and Generative Image Models." arXiv preprint arXiv:2601.08860 (2026).

[10] Shafiq, Muhammad, and Zhaoquan Gu. "Deep residual learning for image recognition: A survey." Applied sciences 12, no. 18 (2022): 8972.

[11] Jiang, Bin, Nanxing Li, Xiaomei Cui, Weihua Liu, Zeqi Yu, and Yongheng Xie. "Research on facial expression recognition algorithm based on lightweight transformer." Information 15, no. 6 (2024): 321.

[12] Xiong, Lingxin, Jicun Zhang, Xiaojia Zheng, and Yuxin Wang. "Context transformer and adaptive method with visual transformer for robust facial expression recognition." Applied Sciences 14, no. 4 (2024): 1535.

[13] Narsimha Reddy, C. H., Shanthi Mahesh, and K. Manjunathachari. "Hybrid feature integration model and adaptive transformer approach for emotion recognition with EEG signals." Computer Methods in Biomechanics and Biomedical Engineering 27, no. 12 (2024): 1610-1632.

[14] Ma, Fuyan, Bin Sun, and Shutao Li. "Facial expression recognition with visual transformers and attentional selective fusion." IEEE Transactions on Affective Computing 14, no. 2 (2021): 1236-1248.

[15] An, Heng-Yu, and Rui-Sheng Jia. "Self-supervised facial expression recognition with fine-grained feature selection." The Visual Computer 40, no. 10 (2024): 7001-7013.

[16] Nawaz, Uzma, Zubair Saeed, and Kamran Atif. "A Novel Transformer-based approach for adult's facial emotion recognition." IEEe Access (2025).

[17] Tagmatova, Zarnigor, Sabina Umirzakova, Alpamis Kutlimuratov, Akmalbek Abdusalomov, and Young Im Cho. "A hyper-attentive multimodal transformer for real-time and robust facial expression recognition." Applied Sciences 15, no. 13 (2025): 7100.

[18] Alzamzami, Fatimah, and Abdulmotaleb El Saddik. "Transformer-based feature fusion approach for multimodal visual sentiment recognition using tweets in the wild." IEEE Access 11 (2023): 47070-47079.

[19] Li, Nianfeng, Yongyuan Huang, Zhenyan Wang, Ziyao Fan, Xinyuan Li, and Zhiguo Xiao. "Enhanced hybrid vision transformer with multi-scale feature integration and patch dropping for facial expression recognition." Sensors 24, no. 13 (2024): 4153.

[20] Liu, Chenyan, and Kezhou Liu. "A Lightweight Transformer Model for Robust EEG Emotion Recognition Using Channel-Wise Differential Entropy." Biomedical Physics & Engineering Express.

[21] Ezati, Ali, Mohammadreza Dezyani, Rajib Rana, Roozbeh Rajabi, and Ahmad Ayatollahi. "A lightweight attention-based deep network via multi-scale feature fusion for multi-view facial expression recognition." arXiv preprint arXiv:2403.14318 (2024).

[22] Siqueira, Henrique, Sven Magg, and Stefan Wermter. "Efficient facial feature learning with wide ensemble-based convolutional neural networks." In Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 04, pp. 5800-5809. 2020.

[23] Singh, Sushil Kumar, Manish Kumar, Ikram Majeed Khan, A. Jayanthiladevi, and Chirag Agarwal. "An Attention-based Model for Recognition of Facial Expressions using CNN-BiLSTM." Polytechnic Journal 15, no. 1 (2025): 4.

[24] Ding, Hui, Shaohua Kevin Zhou, and Rama Chellappa. "Facenet2expnet: Regularizing a deep face recognition net for expression recognition." In 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), pp. 118-126. IEEE, 2017.

[25] Farzaneh, Amir Hossein, and Xiaojun Qi. "Facial expression recognition in the wild via deep attentive center loss." In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 2402-2411. 2021.

Arrived: 17.04.2026
Accepted: 05.07.2026