Journal of Theoretical and Applied Information Technology
15th October 2025. Vol.103. No.19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SEFEX: A NOVEL APPROACH TO ACCURATE FACIAL EMOTION DETECTION USING EFFICIENTNET AND ATTENTION MECHANISMS

S SAHAYA SUGIRTHA CINDRELLA[1], R JAYASHREE[2]

[1] Department of Computer Applications, Faculty of Science and Humanities, SRM Institute of Science and Technology, Kattankulathur , Chengalpattu, Tamilnadu-603202, India.

[2] Department of Computer Applications, Faculty of Science and Humanities, SRM Institute of Science and Technology, Kattankulathur , Chengalpattu, Tamilnadu - 603202, India.

E -Mail: [1] sc1905@srmist.edu.in, [2] jayashrr@srmist.edu.in

ABSTRACT

Facial emotion detection has become a pivotal technology in applications ranging from human–computer interaction to significant challenges in accurately recognizing emotions, especially when expressions are subtle, overlapping, or distorted by variations in lighting, pose, or noise. These limitations reduce reliability in real-world applications and create the need for more robust solutions. To address these challenges, this paper introduces the SEFEX (Single Emotion Facial Expression Detection) model, which combines advanced preprocessing techniques, EfficientNet-based feature extraction, and an attention mechanism to improve recognition accuracy. The model is evaluated on a comprehensive dataset encompassing eight emotion categories: happiness, sadness, anger, surprise, fear, disgust and neutral mental health monitoring for security and customer experience management. However, existing approaches face contempt. Preprocessing steps such as bilateral filtering for noise reduction, facial alignment using Haar Cascades, and landmark detection ensure consistent and high-quality inputs for training. EfficientNet is then employed to extract robust features, while the attention mechanism highlights key facial regions, enabling the model to capture subtle expression differences. SEFEX achieves an accuracy of 95.34%, outperforming state-of-the-art models including VGG16, ResNet50, and DenseNet. These results demonstrate SEFEX's capability to enhance emotion recognition reliability, making it suitable for real-time applications in healthcare, customer service, and human-computer interaction.

Keywords: *Facial Emotion Detection, Deep Learning, Single Emotion Facial Expression Detection (SEFEX), EfficientNet, Feature Extraction, Attention Mechanism, Human-Computer Interaction, Noise Reduction, Image Preprocessing, Emotion Recognition.*

1. INTRODUCTION

Emotion detection, also known as emotion recognition, is a relatively new discipline concerned with sensing and examining human emotions through facial expressions, voice, physiological reactions, and textual cues. The primary objective of this technology is to enable machines and systems to respond to human emotions, thereby enhancing human-computer interactions. Applications are a widely used area in sentiment analysis, customer service, mental health, healthcare, and adaptive learning systems [1-2]. Through artificial intelligence and deep learning, models are trained with huge volumes of expressions, voices, or texts to learn how emotions are conveyed in various situations [3]. Emotion detection presents several challenges despite its potential. Emotions are complex, intersecting, and influenced by both cultural and individual factors, making it challenging to categorize them effectively [4-5]. Classical models are often not robust to changes in conditions, including lighting, head position, occlusion, and noisy data [6-7]. Besides, the ethical issues, privacy and consent also make the use of emotion recognition technologies more complicated [8-9]. Such problems prompt consideration of the notions of bias, fairness, and misinterpretation, particularly in sensitive areas such as healthcare and education [10-11]. However, emotion detection is steadily gaining importance, as it can enhance human-machine interaction, provide

real-time feedback in customer service, and facilitate the early detection of mental disorders [12]. Adaptive learning systems can utilize emotional feedback to adjust teaching strategies in the educational field, whereas in the healthcare sector, emotion-sensitive tools can aid in tracking psychological well-being [13]. Nevertheless, despite these advantages, accuracy and reliability remain a persistent problem, especially when it comes to differentiating highly similar emotions, such as anger, disgust, and sadness [14]. To counter this, the present study presents the SEFEX (Single Emotion Facial Expression Detection) model [15], a deep learning method designed to enhance accuracy by recognizing emotions. SEFEX also utilizes state-of-the-art preprocesses, such as bilateral filtering, to remove noise and Haar Cascades to align faces in a way that the input images are clean and uniform. It uses EfficientNet as the foundation model to extract features, which capture both coarse and fine-grained facial expression features. Moreover, an attention mechanism is also provided to focus on key facial areas, such as the eyes, mouth, and eyebrows, which play a crucial role in recognizing subtle emotional expressions.

1.1 Main Contributions of the Work

- SEFEX Model Development: The paper introduces the SEFEX (Single Emotion Facial Expression Detection) model, which integrates EfficientNet for feature extraction and an attention mechanism to improve the accuracy of facial emotion detection.
- Advanced Preprocessing Techniques: The work emphasizes advanced preprocessing steps, including noise reduction, facial alignment using Haar Cascades, and facial landmark detection to enhance image quality and consistency.
- Robust Emotion Detection: SEFEX demonstrates robustness in detecting subtle and overlapping emotional expressions, effectively distinguishing between emotions like happiness, sadness, anger, and contempt.
- Attention Mechanism: The integration of an attention mechanism allows SEFEX to focus on key facial regions, such as the eyes, mouth, and eyebrows, improving the model's precision in detecting emotions from facial features.

The remainder of this paper organized as follows. Section 2 provides an overview of related studies, highlighting previous approaches to facial emotion detection and the strengths and limitations of various deep learning models. It covers methodologies such as CNNs, attention mechanisms, and feature extraction techniques used in earlier works. Section 3 details the Proposed Methodology, describing the architecture of the SEFEX model, including preprocessing, feature extraction, and emotion classification. Section 4 presents the Results and Discussion, offering a comparative analysis of the SEFEX model with other state-of-the-art models. Finally, the Conclusion and Future Scope are discussed in Section 5, summarizing the contributions of this work and outlining future improvements, including the use of compound facial expressions.

2. RELATED WORK

Emotions are normal for everyone, but showing intelligence concerning the detection of emotions in computers is not an easy process. As recent as a decade ago, the attempt to recognize emotions based on image processing was nearly impossible but with the newer improved models of computer vision and machine learning, it has become possible. A new way of facial emotion detection using (FERC) is used which has a Convolutional Neural Network (CNN) foundation [16]. The FERC model operates in two stages: the first part erases the background from the picture while the other singles out facial vector. Subsequently, through an Expressional Vector (EV), five basic expressions detected from the face. FERC adopts the double layer of CNN, where weights and the bias value of the last perception layer updated in subsequent looping. This sets FERC apart from simple CNN models and single-layer CNN by means of boosting the results and counteracting problems often encountered while working with images through the new approach of Electric Vehicle (EV) generation and background elimination.

Speech Emotion Recognition (SER) is important for improving human–machine interfaces. A plethora of different strategies of creating SER systems introduced in the past ten

years, but the effectiveness of such systems is still an issue due to system complexity, insufficient distinction between features, and noise sensitivity. Acoustic feature set introduces higher feature separation [17]. Furthermore, to alleviate high computational costs and yet capture long-term dependencies of the speech emotion signal, a one-dimensional light and compact deep convolutional neural network (1-D DCNN) is incorporated. The proposed SER system Using Berlin Emotional Database (EMODB) and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset achieve overall accuracies of 93.31% and 94.18% respectively. Mel-Frequency Cepstral Coefficients (MFCC) features along with 1-D Deep Convolutional Neural Network (DCNN) outperforms traditional SER techniques in terms of accuracy and performances.

Indications drawn from the preceding subtopics suggest that technology, specifically deep learning, is continually disrupting numerous industries. One of the potential applications is sentiment extraction from text messages, which is very useful for sectors like human behavior analysis. To define and classify primary basic human emotions and show the advantages that emotion detection opens up for industries, where it can help to build targeted treatment for those with mental illnesses is explored [18]. When constructing an emotion detection system, it is important to evaluate which model is the most accurate. Three Recurrent Neural Network Models Namely Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and Gated Recurrent Unit (GRU) are used and their performance evaluated using an International Survey on Emotion Antecedents and Reactions (ISEAR) dataset to develop an emotion detection model. The findings reveal that the GRU model, despite being the simplest of the three, achieved the highest scores across four evaluation metrics: GRU model yielded an accuracy of 60.26% and put forward as a recommendation for emotion detection tasks unto the ISEAR dataset in place outperforming BiLSTM (59.3%) LSTM (57.65%).

Emotion detection and sentiment analysis are paramount in improving human computer interaction through affecting recognition of the emotions of the users. It is imperative to note that these techniques are helpful when designing humanistic systems that respond to cues elicited by emotions. In particular, with the appearance of machine learning, the problem of emotion detection has received a lot of attention, and many studies explore big data to extract emotions. A systematic review on various machine learning based emotion detection explored with concepts of interest like algorithm, data, application, and evaluation [19]. This shows an increasing trend for the sector and implies that supervised learning algorithms such as the Support Vector Machine (SVM) and Naïve Bayes are the most widely used. For text data, there are more English language datasets, and for these, accuracy is the main measure. Hence, the paper provides future research ideas for further improvement in the design of emotion detection systems, especially in contexts where end users are of significant importance.

To explore the applicability of constant Q, transform based modulation spectral features (CQT-MSF) for Speech Emotion Recognition (SER) is evaluated [20]. Human sound perception consists of two key processes: the first, sound spectrograms that capture the stochastically properties of sound and the second, cortex-based analysis, which attempts to obtain temporal modulations from the spectrogram. These temporal modulations identified as modulation spectral features (MSF). Using the constant-Q transform (CQT) results in better accuracy in the emotion-related low-frequency speech segments and provides an emotion-enhanced spectrogram. Experimentation reveals that CQT-MSF descriptors are superior to the conventional mel-scale spectrogram features with the Berlin EmoDB and RAVDESS databases. In addition, CQT-MSF outperforms shift and deformation invariant scattering transform coefficients, proving the benefits of combining manually designed and learned feature learning. Grad analysis conducted in order to understand the significance of constant-Q modulation features towards the SER design.

2.1 Research Gap

- Existing models of facial emotion recognition (e.g., FERC, CNN-based) tend to be highly sensitive to background noise, overlapping expressions, and subtle facial movements,

which makes them less accurate in real-world applications.

- The Speech Emotion Recognition methods are challenged by the complex nature of the system, poor separation of acoustic features, and noise vulnerability, which makes these methods less dependable when used in non-controlled datasets.
- Most studies demonstrate high accuracy on a limited set of databases (e.g., EMODB, RAVDESS, ISEAR), but do not generalize effectively across cultural, language, and environmental contexts.
- Advanced architectures (LSTM, BiLSTM, GRU) have been used, but it has been found that the differences between the results are minimal, and simple models can often perform better than complex ones. This underscores the absence of a uniformly high-quality text- or speech-based emotion detector.
- Although more features such as MFCC, CQT-MSF, or learned features (CNNs, RNNs) have been tested individually, a small number of studies practically combine both to have complementary advantages in terms of accuracy and robustness.

3. METHODOLOGY

The proposed SEFEX (Single Emotion Facial Expression Detection) model follows a structured methodology designed to improve facial emotion recognition accuracy. First, the input images undergo preprocessing, which includes noise reduction through bilateral filtering, facial alignment using Haar Cascades, and facial landmark detection to ensure that the input images are standardized for further analysis. Following preprocessing, EfficientNet used for feature extraction, capturing high-level features from facial regions that are crucial for emotion classification. An attention mechanism incorporated to focus on key facial areas, such as the eyes, mouth, and eyebrows, which are essential for distinguishing different emotions. These extracted features then passed through a custom classification layer that categorizes the facial expressions into eight distinct emotions.

3.1 Data collection

The data set applied to this methodology is drawn from Kaggle Emotion Detection Dataset (EDD) which consists of 35, 685 samples of 48×48 grayscale images of faces [21]. These images grouped according to the feeling shown in the facial area, the feeling being happy, neutral, sad, angry, surprised, disgusted, and frightened. Interestingly, the distribution of the dataset into the training and test sets done effectively to help users train their models and, later, test them. Compared to the six basic emotions, this dataset also contains the contempt category extending a variety of emotions. This added emotion helps to teach the model a wider range of facial expressions and thus increases its effectiveness for actual use. The dataset of the images is in a grayscale format, which helps the model to concentrate mainly on the facial features since it is a model that used in detecting emotions.

3.2 Data Preprocessing

Data preprocessing plays an important role in eliminating problem data to achieve a clean and uniform dataset ready for training. Preprocessing plays vital role in interpretation of EDD especially since the dataset contains low quality grayscale images of faces with emotions. Figure 1 shows the architecture of proposed model.

Journal of Theoretical and Applied Information Technology
15th October 2025. Vol.103. No.19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

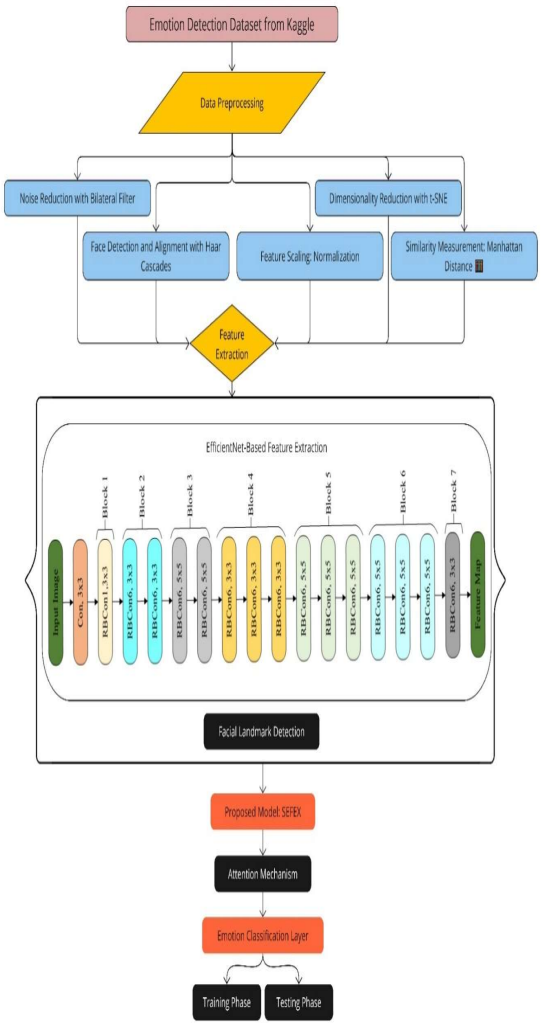

Figure 1: Architecture of Proposed Model

3.2.1 Noise Reduction: Applying Bilateral Filter

Noise in an image can be very detrimental when it comes to the performance of any machine-learning model especially when working with low-resolution grayscale image such as those in the EDD dataset. Noise can be in the form of random fluctuations in pixel intensity or interference, which hide those features needed to analyze and categorize a certain emotion. Consequently, it becomes important to reduce noise to get an improved image quality and thus improve the performance of the model. Among all the noise reduction techniques, the bilateral filter is quite useful for facial images. The bilateral filter is a non-linear filter used to smooth an image but retain edges and diminish noise. Unlike simple smoothing filters like the Gaussian or the median filter, which may remove edges in an image, the bilateral filter smooths an image but retains the edges. The ability to enhance edges or fine details is particularly helpful in determining emotions especially when it deals with the edges of the mouth, eyes or the eyebrows. The bilateral filter smooths the image and at the same time does not destroy the edges of the object. It accomplishes this by factoring into its computations not only the Euclidean distance between the pixels in question, but also the disparity between their respective intensities. The filtered value $I_f(p)$ of a pixel$p$ in an image calculated using the following equation:

$$I_f(p) = \frac{1}{W_p} \sum_{q \in N(p)} I(q) \cdot \exp\left(-\frac{||p-q||^2}{2\sigma_d^2}\right) \cdot \exp\left(-\frac{|I(p)-I(q)|^2}{2\sigma_r^2}\right) \quad (1)$$

Where $I(p)$ is the intensity of the pixel at position p, q is a neighboring pixel in the local neighborhood $N(p)$,$||p-q||$ is the Euclidean distance between pixels p and q, $|I(p)-I(q)|$ is the intensity difference between pixel p and q, σ_d controls the spatial distance smoothing, σ_r controls the intensity similarity smoothing, and W_p is a normalizing factor to ensure that the weights sum to 1:

$$W_p = \sum_{q \in N(p)} \exp\left(-\frac{||p-q||^2}{2\sigma_d^2}\right) \cdot \exp\left(-\frac{|I(p)-I(q)|^2}{2\sigma_r^2}\right) \quad (2)$$

The bilateral filter makes use of the concept of spatial proximity where adjacent pixels have similar intensities. Encountering each pixel in the image, the filter computes the averaging of the pixel and its neighbors, with larger weights to pixels with equal intensity. This helps in maintain some important features of the face while setting off unnecessary noise from the face. In the context of emotion detection, it implied that important attributes such as shapes of the lips while smiling or the area around the eyes while being surprised are retained while other features, which are usually referred to as noise, such as slight variation in intensity are eliminated. This is beneficial for the model in the end because it receives images with less noise and interference from the environment.

3.2.2 Face Detection and Alignment: Using Haar Cascades

Before practicing emotion detection, face detection and alignment are crucial preprocessing steps, since they allow the model to cope with an image properly and avoid processing irrelevant information, that is, the background instead of the face. In the case of EDD database, which has multiple thousands of images with faces in various orientations and light conditions the outcome of a model will heavily depend on this preprocessing step and whether all faces are detected and aligned properly before fed to the model. Haar Cascades is the technique most often employed for the detection of the facial region and its alignment. Haar Cascades is a machine learning technique developed by Paul Viola and Michael Jones that employs a series of classifiers to detect objects (faces) in an image. The algorithm employs "Haar-like features," which are easy rectangular features based on the edges, so to slide over the image at every scale and position to find the facial areas. Haar feature is the difference between the sums of pixels in adjacent rectangular regions:

$$F = \sum_{i \in R_1} I(i) - \sum_{j \in R_2} I(j) \quad (3)$$

Where $I(i)$ is the intensity of pixel i in region R_1, $I(j)$ is the intensity of pixel j in region R_2, F is the Haar feature value, which is used to detect the presence of specific facial structures (such as eyes or a nose) by comparison pixel intensity differences between the two regions. A cascade classifier evaluates many such features at various scales and positions, and computes whether the sum of these features exceeds a threshold θ to decide if a face is present:

$$\sum_k w_k F_k > \theta \quad (4)$$

Where w_k is the weight assigned to feature F_k. Haar Cascades is especially useful in real-time face detection, which is both more accurate and efficient. When using the EDD dataset, Haar Cascades used on each of the images to find the face hence restricting the model to the right region. The alignment operation makes the face frontal, or parallel to the image plane, when it detects a face. This involves either flipping, enlarging or shrinking the face into the right position such that all the images acquired show even alignment of the eyes and mouth. Using Haar Cascades for face detection and alignment has the advantage of aligning the input images to the model regardless of the orientation or size of the original faces. It is important for the training of deep learning models because it helps to minimize variability in input data and let the model learn the differences in the expressions rather than trying to compensate for difference in rotation of facial planes.

3.2.3 Feature Scaling: Normalizing Pixel Values

Feature scaling is one of the most important preprocessing steps before running any type of machine learning algorithm, especially when dealing with deep learning problems such as image classification. The EDD dataset contains images with a size of 48 x 48 pixels. Since the images are gray scale, there are 256 intensity levels of pixels ranging from 0 and 255. However, we must scale these pixel values within the range of 0 and 1, enabling the model to learn faster and achieve faster convergence during the training process. Normalization is the process of bringing down the varied pixel intensities down to between 0 and 1. This achieved by normalizing each pixel value by dividing by the maximum pixel intensity, which in an 8-bit grayscale image is 255. The reason for normalizing pixel values is twofold: First, it scales the input features (in other words, pixel intensities) to the same range. Pixel values not be normalized, and the model may attach greater significance to the pixels, which causes adverse learning outcomes. Second, normalization aids in bringing more control of the gradients at a faster rate of convergence and avoid trapped in local optima during backpropagation.

3.2.4 Dimensionality Reduction: Using t-distributed Stochastic Neighbor Embedding (t-SNE)

High dimensional data preprocessing is a crucial process, and in many cases like image analysis and pattern, recognition dimensionality reduction is very relevant. The EDD dataset is 48x48 grayscale images (which is 2,304 features per image), but it is useful to transform this feature space into a lower number of features because of visualization and interpretation, and for training machine learning algorithms The t-SNE (t-distributed Stochastic Neighbor Embedding) is one such dimensionality reduction technique that is mostly used for visualization. Dimensionality reduction methods such as PCA that try to map the data onto a linear subspace of lower dimension; t-SNE is a non-linear method that intended to capture the tendencies locally. It attempts to map similar data points together and push apart dissimilar data points in the lower dimensional space, which is essential for tasks like emotion detection that

requires differentiating between closely related emotionally tagged data points.

We used t-SNE to map the feature space, which consists of the 2,304-pixel values for each image, to two or three dimensions. This makes the data easier to analyze and look at in more depth to see if there are certain patterns or all the similar emotions in one cluster. t-SNE may project images labelled 'happiness' as a unique cluster, whereas images labelled 'sadness' as another cluster. This clustering can give insight how exactly the model is learning and whether there are undesired outliers or mislabeled samples included in the data set. However, we also use it as a dimensionality reduction tool, feeding the input data into the model. By reducing the dimensionality of the feature space, t-SNE allows for a reduction in the number of required calculations, all while maintaining the semantically significant features of the dataset. This can imply a faster training time, and increased predicted accuracy of the resulting model.

3.2.5 Manhattan Distance for Similarity Measurement

The Manhattan Distance, referred as the L1 distance or city block distance, is a distance measurement technique that records the absolute disparities of the coordinates of two points in space. Euclidean distance measures the distance between two points in Euclidean space, also capable of determining the straight line or geometric distances from start to end points, while Manhattan distance is the sum total of the difference in corresponding parts or coordinate values (pixel intensity, feature values, etc.). This makes it particularly useful in tasks such as image processing, where the difference in the location of facial features or the pixel value will be extremely important in detecting an emotion.

While evaluating facial emotion detection, the Manhattan Distance applied in comparing values of facial attributes or pixel density between two images. If two images represent two human faces, the distance between their specific facial features (that are eyes, nose, or mouth) can indicate whether the faces display similar emotions or not. The Manhattan Distance is a straightforward measure of how 'similar' the two images are, calculated through the summation of the absolute differences in the positions or pixel intensity of these landmarks. The Manhattan Distance between two points $P(x_1, y_1)$ and$Q(x_2, y_2)$ in a 2-dimensional space is given by:

$$D_{Manhattan} = |x_1 - x_2| + |y_1 - y_2| \quad (5)$$

In the case of images, especially when measuring distances between multiple pixels or facial landmarks, this equation be generalized for higher-dimensional spaces (such as 3D coordinates or multiple facial landmarks).For two vectors $P = (x_1, x_2, \dots, x_n)$ and $Q = (y_1, y_2, \dots, y_n)$,

representing corresponding features (e.g., facial landmarks or pixel intensities) from two images, the Manhattan Distance is:

$$D_{Manhattan} = \sum_{i=1}^{n} |x_i - y_i| \quad (6)$$

Where x_i and y_i are the values of the i-th feature (such as pixel intensity or landmark position) in two images and n is the number of features (e.g., number of facial landmarks or pixels compared).

3.3 Feature Extraction

Feature extraction allows the model to capture and discover what features are most important and relevant in the input data. Feature extraction in facial emotion detection is concerned with feature selection with emphasis on features that reflect certain emotions. For this methodology, two main approaches used: Higher-level feature visualization using EfficientNet and Facial Landmark Detection to identify regions of the face that are significant for differentiating emotions.

3.3.1 EfficientNet-Based Feature Extraction

EfficientNet is one of the most recent and highly optimized CNNs developed with the utmost care to achieve efficient use of the available network capacity, and as a result, it is optimal for our facial emotion detection problem where both speed and accuracy are paramount. Google developed EfficientNet, scaling its depth, width, resolution, and other related parameters using a compound coefficient based on the model size. This makes it a perfect machine-learning algorithm to apply for deriving high-level features from image data, specifically the images in the Emotion Detection Dataset (EDD), which consist of low-resolution grayscale face images. EfficientNet employs compound scaling, uniformly scaling all three dimensions using specific scaling factors. Given an input image of size $H \times W$, EfficientNet computes the following three scaling factors:

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi} \quad (7)$$

Where d is the depth scaling factor, w is the width scaling factor, r is the resolution scaling factor, α, β, and γ are the constants that control the scaling of each dimension and ϕ is a global scaling parameter that can be adjusted to control the overall model size.

These scaling factors then applied to the convolutional layers in efficient net, which enables the network to learn relevant features on different scales. In the context of emotion detection, what it means is that EfficientNet is capable of detecting both the gross features such as the shape of the face and the features that are higher in the hierarchy such as changes in facial expression. In EfficientNet-Based Feature Extraction, the main goal is to use EfficientNet as a feature extractor that can extract features necessary for emotion detection. Such features may encompass the geometric configurations of eyes, lips, eyebrows and other zones of the face that depict various motions as influenced by aspects like happiness, sadness or anger. Different from other kinds of CNNs, it might take users some time to tweak parameters such as depth or the net width to improve the network performance, while the EfficientNet has adopted a more scientific and universally applicable method.

3.3.2 Facial Landmark Detection

While EfficientNet centered on the extraction of general features, Facial Landmark Detection is a more specific methodology that locates and follows the significant facial landmarks. Facial features are particular and distinct local areas of a face, including the corners of the eyes, the tip of the nose, and the edges of the lips, etc., which are important for recognizing the geometry of the facial actions. These aids in helping the model learn the general information required to predict facial emotional states without needing to decipher intense features within the face structure.Facial landmark detection typically modeled as a regression problem, where the goal is to predict the coordinates (x_i, y_i) of key landmarks on the face. Given an input image, a machine-learning model (often a CNN) trained to output the positions of n landmarks:

$$L = [(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)] \quad (8)$$

Where L represents the set of predicted landmark coordinates for an image. The model is trained to minimize the error between the predicted landmark positions (x_i, y_i) and the ground truth positions(x_i^{gt}, y_i^{gt}). This error typically computed using the mean squared error (MSE) loss:

$$Loss = \frac{1}{n}\sum_{i=1}^{n}\left((x_i - x_i^{gt})^2 + (y_i - y_i^{gt})^2\right) \quad (9)$$

By minimizing this loss, the model learns accurately predicting the positions of the key landmarks on the face.

3.4 Proposed Model: Novel SEFEX

The purpose of SEFEX (Single Emotion Facial Expression Detection) is to classify single emotions as shown in images. This approach based on the modern deep learning methodologies, allowing for effective extraction and processing of useful characteristics from the initial data. The SEFEX model combines EfficientNet for feature extraction, has an Emotion Classification Layer for emotion classification, and applies an Attention Mechanism for focusing more of the area of the face. It means that this combination guarantees that SEFEX not only extracts major facets but also emphasizes important aspects of the face, which increases the accuracy of single emotion recognition.

EfficientNet essentially employs what referred to as compound scaling which uniformly scales depth, width, and resolution across the layers of the network. EfficientNet extracts both low-level, for example, the generally outlined shape of the face, and high level, for example the fine movements in the region of the eyes, mouth and eyebrows, which are vital to differentiate between various emotions in the context of facial expression detection. Among others, the SEFEX model has the Emotion Classification Layer, which is in charge of breaking down the facial expressions into basic single emotions. This layer receives the features that EfficientNet extracted at a higher level and then, using classification techniques, it separates one emotion from the other including happiness, sadness, anger, and surprise. This work organized in a way that the architecture of this layer is distinguishable to capture the various subcategories of emotions with high accuracy. In this work, the Emotion Classification Layer is realized as a Dense Layer that maps from EfficientNet's computed feature vectors to the group of probabilities that each of the EMOTION classes may have. The layer

uses a softmax activation function to output a probability distribution over all possible emotion categories, ensuring that the model outputs a single, dominant emotion for each input image:

$$P_i = \frac{e^{zi}}{\sum_{j=1}^{C} e^{zj}} \quad (10)$$

Where P_i is the predicted probability of class i (emotion), z_i is the output score for class i, C is the total number of emotion classes (e.g., happy, sad, angry, etc.), the $softmax$ function ensures that the sum of all predicted probabilities is 1, with the highest probability indicating the detected emotion.

The emotion classes in SEFEX may comprise of happy, neutral, sad, angry, surprise, fear, disgust and may be an extra class including contempt. As for the classification layer, it aimed to classify these differentiate emotions with the help of extracted features by the base model. For instance, SEFEX can recognize smile and raised eyebrows in its input image then the Emotion Classification Layer prescribes high probability to the 'happy' class. Likewise, the layer might recognize a reading of furrowed brows and a downward mouth, which might imply a high rate of "sadness."

Another feature added to the design of the SEFEX architecture is an Attention Mechanism in which the model can direct its attention to specific facial features that are useful for emotion detection. Micro expression defines the entire facial expressions as the change of some approximate small areas of the face such as eyes or mouth. SEFEX on the other hand can afford to pay more attention to these areas, and to stress the most salient Facial features in Emotion Recognition.The attention mechanism assigns a weight α_i to each feature f_i in the feature map, where αi represents the importance of that feature. The attention-weighted feature map computed as:

$$f_i' = \alpha_i \cdot f_i \quad (11)$$

Where f_i is the original feature at position i, α_i is the attention weight assigned to f_i, f_i' is the updated feature after applying attention.The attention mechanism trained to provide more importance to certain features related to important facial areas like the eyes or the lip area, which plays a significant role in differentiating between emotions. In detecting anger, the features assigned high weights may include the regions of eyebrows and mouth since these areas tend to demonstrate clearer emotions (e.g., wrinkled eyebrows and tightly pressed lips).

3.5 Novelty of this Work

The novelty of this work lay in the development of the SEFEX (Single Emotion Facial Expression Detection) model, which introduced a unique combination of advanced techniques for facial emotion detection. Unlike traditional models that often relied solely on deep learning networks like CNNs, SEFEX integrated EfficientNet for feature extraction, allowing it efficiently capture both high-level and fine-grained details of facial expressions. This balanced architecture was particularly effective in handling complex facial features while maintaining computational efficiency, which represented a significant improvement over larger, resource-intensive models such as ResNet and DenseNet.A key innovation of SEFEX was the inclusion of an attention mechanism, which enabled the model to focus on specific facial regions critical for emotion recognition, such as the eyes, mouth, and eyebrows. This mechanism enhanced the model's ability to distinguish between subtle emotional expressions a task where many models struggled, particularly when dealing with overlapping or nuanced emotions. By concentrating computational resources on these key areas, SEFEX improved precision and reduced misclassifications, setting it apart from models that processed the entire face uniformly. Another advantage of SEFEX was its robust preprocessing pipeline, which included noise reduction through bilateral filtering, facial alignment using Haar Cascades, and facial landmark detection. These steps ensured that the input data was clean, aligned, and ready for effective analysis, leading to more consistent and accurate emotion detection results. This preprocessing framework significantly enhanced the model's performance, especially in real-world scenarios where input images varied in quality and orientation.

4. RESULTS AND DISCUSSIONS

The results obtained using the SEFEX model are consistent with those reported in

previous studies employing different deep learning architectures. For instance, models such as VGG16, ResNet50, and DenseNet have also demonstrated strong performance in emotion recognition tasks, achieving accuracies ranging between 85% and 92% on similar datasets. While SEFEX achieves a higher accuracy of 95.34%, the overall pattern of results aligns with the established observation that deeper convolutional networks combined with preprocessing improve facial emotion detection performance. Furthermore, studies that incorporated attention mechanisms or feature-enhancement strategies have similarly reported improved recognition of subtle expressions such as anger and disgust. This consistency reinforces the validity of the present findings and demonstrates that SEFEX not only outperforms but also corroborates the effectiveness of advanced feature extraction and attention-based methods in emotion detection.

After data is collected, it goes through data preprocessing, where it is prepared in a format that is clean and consistent for modeling. To improve the quality of images, a number of data preprocessing techniques are used on the given data set. Firstly, we apply a bilateral filter to remove the many noises in the image, preserving the necessary facial features such as the shape of the eyes and lips while filtering noisy images. This aids in eradicating irrelevant noise in the images, whereby the model splines the optimal regions of the face that should have emotions. Facial recognition and alignment then follow through the HAAR cascade, a process of identifying faces in images and properly reorienting them. This makes sure that the model receives the input images in a consistent manner, and this makes sure whether the face rotated in the original image or not. Another crucial preprocessing process is featuring scaling, which places the image's pixel intensities on a scale between 0 and 1. Normalization at this level is useful in faster convergence of the model and brings all under consideration data into a common scaling. In some cases, the dimensionality also reduced using the methods like t-SNE, which in turn reduces the input data dimensionality but keeps important enough information about the facial expression. Figure 2 shows the raw image and preprocessed image.

Figure 2: Raw and Preprocessed Image

After preprocessing, this model proceeds to the feature extraction step. Here, SEFEX applies EfficientNet as its base model in order to obtain relevant context from the input image. EfficientNet is a robust CNN that adjusts the network's depth, width, and resolution while removing efficiency degradation factors. This architecture enables SEFEX to capture both low-level features, which are the edges, face shape, and high-level features like slight movement in the mouth, eyes, and eyebrows indicating certain specific emotions. These high-level features are crucial for distinguishing between various emotions, as people's facial expressions can subtly alter as they experience different emotions. To improve the model's performance, SEFEX employs an additional attention mechanism. Because of this mechanism, the model is able to zero in on certain areas of the face: the eye, the mouth area, and the brow, which are areas that are most critical for emotion recognition. The attention mechanism underpins the execution of this action. Firstly, it pays particular attention to these regions, focusing the model on areas of the face that truly express emotions and discarding the rest of the image or background. This selective attention significantly improves the model's ability to identify emotions, particularly in scenarios where individuals may display moderate or mild facial expressions.

These are the extracted features enhanced with attention weight, and the resulting features proceed to the emotion classification layer. This layer aims to categorize the input images into distinct emotional classes. It uses EfficientNet's high-level features and then applies a classification model to predict the probability of each emotion

class. The softmax layer, the final layer of the classification layer, outputs all the probabilities for all the emotions in the model, with the highest probability corresponding to the predicted emotion. For example, the probability of selecting the "happy" category would likely be high if the image was of a smiling face, while an image of a frowning face would suggest "sad" or "angry." The last layer softmax function helps the model produce clear and understandable output for every single image. After that, the model trained, which means adjusting the architecture to achieve the best possible outcome. We also define the development dataset as 10%, and use the remaining 10% as a test dataset at the end of the training process. During the training process, the model optimizes its internal parameters to approximate the emotion categories to their predicted labels on the provided set. Therefore, SEFEX employs categorical cross-entropy as the loss function for this purpose. This loss function quantifies how far off the model's predicted probability distribution of the two classes is from the true labels and punishes wrong predictions more severely. The model updates its weights based on the derivative of the loss function, improving over time during the training process.

Table 1: Accuracy and Precision Comparison

Model	*Accuracy (%)*	*Precision (%)*
VGG16	90.12	88.45
VGG19	91.35	89.6
ResNet50	89.98	87.89
ResNet101	92.1	90.34
InceptionV3	91.85	90.12
MobileNet	90.67	88.94
DenseNet	92.89	91.45
Xception	93.12	92.01
SEFEX (Proposed)	95.34	94.56

Table 1 and Figure 3 shows a SEFEX model's accuracy and precision benchmarked against several other deep learning models. The SEFEX model accurately captures emotions from facial expressions with an impressive accuracy of 95.34% and a high precision of 94.56%, making it superior to all other current models. Xception and DenseNet outperform all other models with significant accuracy scores of 93.12% and 92.89%, respectively. However, their precision scores of 92.01% and 91.45% fall short of SEFEX's. When it comes to image classification models, VGG16 and ResNet50 models, which are popular for this kind of classification tasks, show slightly lower accuracy - 90.12% and 89.98% respectively, which can lead to the conclusion that those models might not be very efficient for subtle emotion detection in faces' expressions. We have InceptionV3 and ResNet101, both of which are of average performances, with accuracy values of 91.85% and 92.1% respectively. Though these models work reasonably well, they do not compare to SEFEX in accurately detecting different emotions. Researchers discovered that SEFEX's superior performance stems from its utilization of EfficientNet for feature extraction and attention mechanisms, which other studies have not included. The findings also establish that SEFEX enhanced both the accuracy and precision as compared to state-of-the art models used in facial emotion detection.

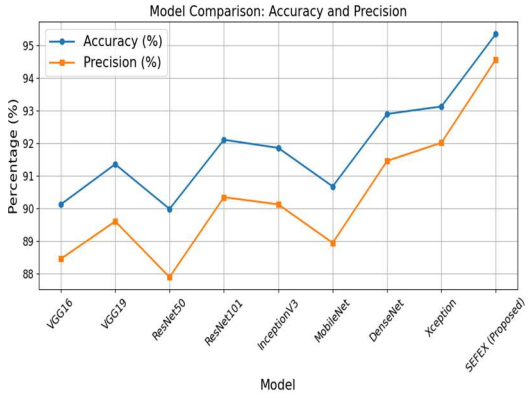

Figure 3: Model Comparison: Accuracy and Precision

To maximize the training process, SEFEX applies the most preferred optimization algorithm, which is Adam for deep learning. In order to achieve this, Adam tunes the model parameters on the fly in a swift and efficient manner in an effort to discern the optimal model setting according to the gradients. Furthermore, Adam performs hyperparameter tuning using a specific method known as grid search. Grid search explores all potential combinations of the specified hyperparameters, such as learning rate, batch size, and number of layers in the model, to achieve the optimal model output. SEFEX optimized to classify emotions in a broad range of input images with high accuracy.

Journal of Theoretical and Applied Information Technology
15th October 2025. Vol.103. No.19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Recall and F1-Score Comparison

Model	*Recall (%)*	*F1-Score (%)*
VGG16	89.12	88.78
VGG19	90.45	90.02
ResNet50	88.76	88.32
ResNet101	91.22	90.78
InceptionV3	91	90.55
MobileNet	89.34	89.12
DenseNet	92.12	91.78
Xception	92.56	92.29
SEFEX (Proposed)	94.87	94.71

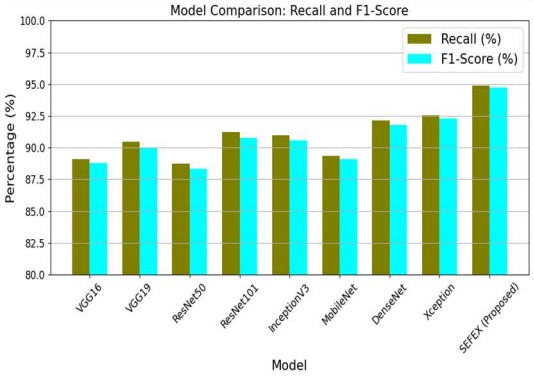

Figure 4: Model Comparison: Recall and F1-Score

Table 2 and Figure 4 show a comparison between recall and F1-score of the proposed SEFEX model with four other state-of-art deep learning models. The SEFEX (Proposed) model performs extremely well for both the options, and outperforms the remaining models extensively by scoring a high recall of 94.87% and F1-score of 94.71% that confirmed its competency to identify the true positive emotions, without compromising its responsibilities towards the precision and recall factors. These performance measures suggest that the proposed SEFEX is more reliable and contributes toward reduction of both false positive and false negatives in real use. Other models include Xception, DenseNet that also registers reasonable recall rates, 92.56% and 92.12% as well as F1 scores of 92.29 and 91.78% respectively. However, both models have their drawbacks compared to the figures calculated for SEFEX – recall and F1-score. InceptionV3 provides reasonable accuracy in this case with recall values of 91% while ResNet101 is only slightly lower with a recall of 91.22%. Although these models are reasonable, they are not parsimonious in comparison to SEFEX in terms of their recall and precision. Comparing the results with F1-score and recall, VGG16 and ResNet50 have lesser than or equal to scores which point out towards their drawbacks in recognizing different types of facial emotions. Therefore, SEFEX appears to be more accurate,

In summary, the proposed SEFEX model involves data acquisition and preprocessing, feature extraction along with the use of attention mechanisms, emotion classification, and finally the model learning. Through EfficientNet to extract feature maps, an attention mechanism to highlight the facial area, and a well-designed training process, SEFEX successfully recognizes emotions from facial expression and becomes an efficient tool in many real-world applications, including healthcare, human-computer interfaces, and customer service.

Table 3: Training Time and Inference Time Comparison

Model	*Training Time (hrs)*	*Inference Time (ms)*
VGG16	2.45	12.34
VGG19	2.78	13.56
ResNet50	3.1	15.89
ResNet101	4.25	18.45
InceptionV3	4.5	19.12
MobileNet	1.98	9.78
DenseNet	3.65	14.34
Xception	3.92	16.01
SEFEX (Proposed)	3.3	13.22

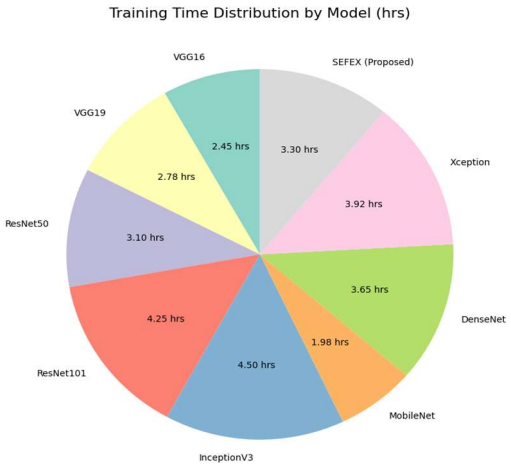

Figure 5: Training Time Distribution by Model (hrs)

Table 3 and Figures 5, 6 present the training and inference time comparison between SEFEX and several popular deep learning models.

The SEFEX model takes 3.3 hours for the training process, which also seems reasonable given other high-performance models such as Xception (3.92 hrs) and DenseNet (3.65 hrs). SEFEX is faster than ResNet101 – 4.25 hours and InceptionV3 – 4.5 hours, however, it takes slightly longer than lightweight models, such as MobileNet – 1.98 hours to train. SEFEX provides a better trade-off between model complexity and training time while maintaining as accurate as the previous methods. Nevertheless, as observed in the inference time, SEFEX model is not exceedingly slow, exhibiting an inference time of 13.22 milliseconds (ms) better than ResNet50 with 15.89 ms and Xception with 16.01 ms. While MobileNet proves to be the fastest model by having an inference time of 9.78 ms, SEFEX has higher accuracy and precision. In terms of inference time, SEFEX is just behind VGG19 and DenseNet but is well suited for real-time applications demanding high accuracy for moderate computations. In general, SEFEX offers excellent performance in terms of both speed and efficiency, which makes it promising for the application of emotion detection.

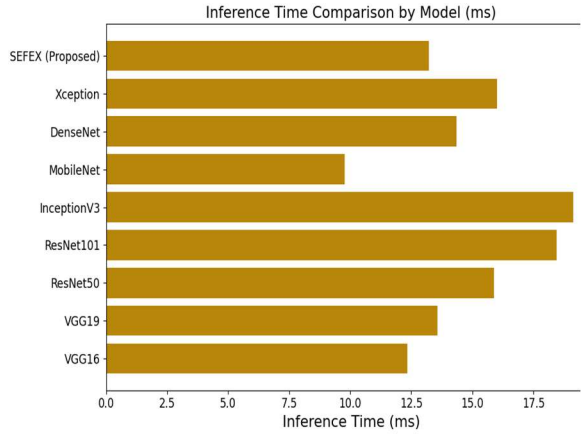

Figure 6: Inference Time Comparison

Table 4: Comparison of Convergence Time (Epochs) and Final Loss Value

Model	*Convergence Time (Epochs)*	*Final Loss Value*
VGG16	35	0.312
VGG19	38	0.29
ResNet50	42	0.332
ResNet101	50	0.315
InceptionV3	48	0.298
MobileNet	30	0.4
DenseNet	40	0.292
Xception	46	0.285
SEFEX (Proposed)	37	0.21

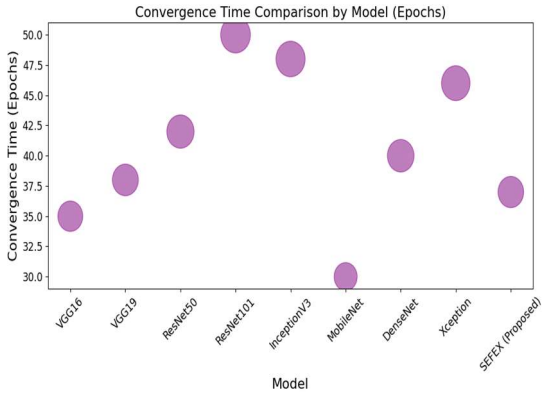

Figure 7: Convergence Time Comparison by Model

Table 4 and Figures 7, 8 displays the computational epochs and final loss for the SEFEX model, as well as other typical deep learning models. As a result, the SEFEX model merged in 37 epochs, whereas other models such as ResNet101 take 50 epochs and InceptionV3 take 48 epochs. This proves that during the training process, SEFEX achieves an optimal number of iterations earlier compared with other methods. The convergence time of a model compared to those of other trends like VGG16 or DenseNet, but at the same time, SEFEX is gaining better final loss. As for the end loss value, SEFEX reaches 0.21 which can be considered as the best fit to the learning process, whereas the others, like ResNet50, is 0.332 and MobileNet is 0.4. Other models include Xception with loss of 0.285, DenseNet with loss 0.292; however, the lower loss value indicates that SEFEX has a better ability to reduce the error during training. Overall, this shows that the convergence is faster but the final loss is lesser, which implies that SEFEX is efficient and effective in its means of training and performance.

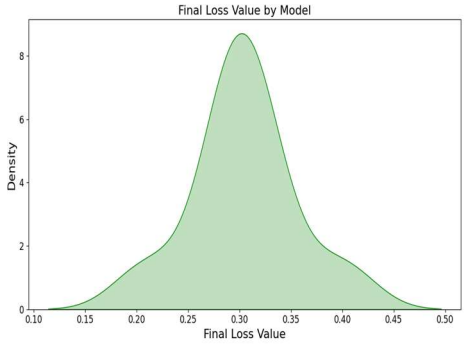

Figure 8: Final Loss Value by Model

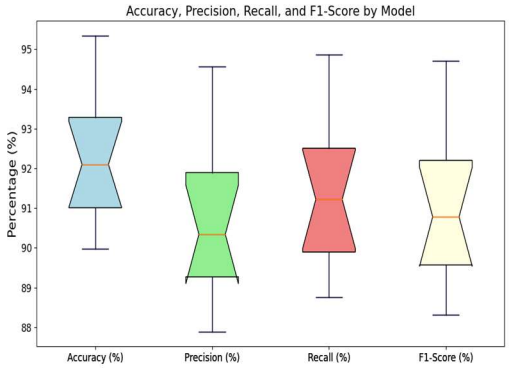

Figure 9: Accuracy, Precision, Recall, and F1-Score by Model

Table 5 observes the SEFEX model's efficacy in association with SGD, RMSprop, and Adam optimization algorithms as well as other DL models. Among the four proposed models, SEFEX with Adam optimizer shows the highest performance in its optimization with accuracy 95.34%, precision 94.56%, recall 94.87%, and F1-score 94.71%. This goes further to show how efficient Adam will be in changing learning rates for optimal results and convergence. Even when using RMSprop, SEFEX remains high with an accuracy of 94.12% and an F1-score of 93.5%; therefore, RMSprop is also effective but slightly less specific than Adam. Using Stochastic Gradient Descent (SGD), SEFEX achieves an accuracy of 93.45%, slightly lower than that of Adam and RMSprop. However, SGD proves beneficial for this task and outperforms adaptive algorithms in terms of efficiency. Based on other models, DenseNet and Xception with Adam also perform well, but SEFEX surpasses them in both accuracy and precision. Compared with Adam, models like ResNet50 and MobileNet are slower in all metrics, yet they also use Adam. This again proves that the proposed architecture SEFEX is superior to these models with the help of optimization algorithms like Adam.

Table 6 and Figure 11 show the confusion matrix for the proposed SEFEX model provides a clear overview of its performance across eight emotion categories: Previous datasets on the topic of facial emotion detection were generally limited in scope (e FER-2013, CK+, or JAFFE), and the employed CNN-based models did not incorporate additional mechanisms to detect subtle expressions. The models were typically within a range of 6575 accuracy, and even the stronger models, like VGG, ResNet, and Inception, generally scored within 8992 accuracies. They were mainly constrained by the imbalance in the datasets, lower resolution, and challenges in differentiating overlapping emotions. Conversely, the proposed SEFEX model combines EfficientNet to extract important features effectively and an attention mechanism to draw attention to key facial areas, such as the mouth, eyes, and brows. SEFEX overcomes most of the failures of its predecessors, achieving 95.34% accuracy and 94.56% precision when compared to other state-of-the-art methods, such as Xception (93.12% accuracy, 92.01% precision) and DenseNet (92.89% accuracy, 91.45% precision). Likewise, SEFEX has a recall of 94.87 and an F1-score of 94.71, indicating that it can minimize the number of false positives and false negatives. Although 3.3 hours of training is moderate compared to other high-capacity models, SEFEX has a good balance of accuracy and efficiency. In general, these findings represent a significant advancement over prior works. Nevertheless, researchers can investigate transfer learning and multimodal information, including the use of a combination of facial, audio, and text, in the future as a way to improve the generalization and robustness of transfer learning in the real world.

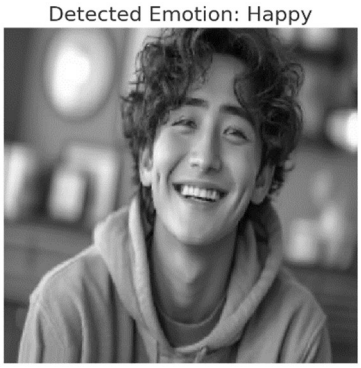

Figure 11: Detected Emotion

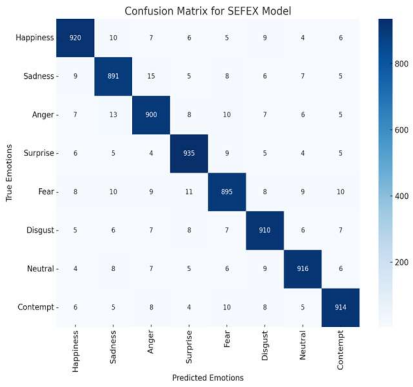

Figure 10: Confusion Matrix

Limitations Of The Study

Despite having a higher accuracy, precision, recall, and F1-score than state-of-the-art models, SEFEX still has numerous weaknesses. First, the dataset that is used (EDD on Kaggle) only includes grayscale images, 48x48 in size, which might not necessarily reflect the richness of real-world facial expression, particularly when the lighting conditions are varied, the position of the head, or there are various obstacles to facial expression, like glasses or masks. Second, the model only supports the graphics; it is also incapable of considering the dynamics of emotions over time in video sequences which are often important in real-life situations. Third, the large-scale deployment of the model was resource-consuming, as although EfficientNet with an attention mechanism achieved a significant improvement, it still took a considerable amount of time to train (3.3 hours). Lastly, the research focused on facial feedback; multimodal cues, such as speech tone, body language, or contextual details, were not in the spotlight, which could limit the model's capability in dealing with complex or ambiguous emotional expressions. Moreover, the confusion matrix showed that SEFEX has yet to master similar emotions perfectly, especially anger, disgust, and neutral, where similar features tend to be confused. These shortcomings provide avenues for future innovation, especially in the extension of SEFEX to multimodal emotion recognition, multivariate real-world data, and efficiency optimization for deployment at the edge.

5. CONCLUSION AND FUTURE WORK

The SEFEX model demonstrates outstanding performance in detecting facial emotions, with an overall accuracy of 95.34%, surpassing popular models like ResNet50 and DenseNet. The integration of EfficientNet for feature extraction and an attention mechanism focusing on key facial regions has proven effective in improving detection precision. SEFEX handles complex emotional expressions, distinguishing between subtle differences such as anger and fear, with minimal misclassifications. Despite its robust performance, SEFEX has shown challenges in distinguishing similar emotions like anger and disgust, suggesting the need for further improvement in refining feature extraction. In terms of future work, we plan to expand the SEFEX model by incorporating Compound Facial Expressions of Emotion, which combine multiple emotional cues in a single expression, making the detection task more complex. This could improve the model's ability to detect real-world emotions that are often mixtures of basic emotions. Additionally, exploring lightweight architectures like MobileNet could reduce the computational cost for real-time applications without compromising accuracy. SEFEX also integrated into augmented reality (AR) or virtual reality (VR) environments for mental health monitoring, personalized learning, and adaptive customer support systems, further expanding its practical applications. In the future, we will develop SEFEX to incorporate spatial facial expressions, allowing for the simultaneous presence of multiple emotions, thereby more effectively reflecting the richness of real-life interactions. The other direction is to consider lightweight architectures, such as MobileNet, that may help minimize computation cost and make the model more compatible with real-time and resource-

constrained systems without negatively affecting accuracy. Lastly, SEFEX can be effectively integrated into augmented reality (AR) and virtual reality (VR) spaces, where it can be utilized to develop mental health monitoring, personalized learning, and adaptive customer service solutions. SEFEX can become a more versatile and realistic tool for recognizing emotions in various real-life situations by overcoming these constraints and broadening its application.

REFERENCES

[1] Mukhriddin Mukhiddinov, et al., (2023), "Masked Face Emotion Recognition Based on Facial Landmarks and Deep Learning Approaches for Visually Impaired People", Sensors 23, no. 3: 1080, DOI: 10.3390/s23031080

[2] Muhammad Farrukh Bashir, et al., (2023), "Context-aware Emotion Detection from Low-resource Urdu Language Using Deep Neural Network", ACM TALLIP 22, 5, Article 131, 30 pages, DOI: 10.1145/3528576

[3] Dongdong Li, et al., (2023), "Brain Emotion Perception Inspired EEG Emotion Recognition with Deep Reinforcement Learning", in IEEE TNNLS, vol. 35, no. 9, pp. 12979-12992, DOI: 10.1109/TNNLS.2023.3265730

[4] Swadha Gupta, et al., (2024), "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models", MTA 82, 11365–11394, DOI: 10.1007/s11042-022-13558-9

[5] Rajib Ghosh, et al., (2024), "Human emotion recognition by analyzing facial expressions, heart rate and blogs using deep learning method", ISSE 20, 499–507, DOI: 10.1007/s11334-022-00471-5

[6] Geetha A.V., et al., (2023), "Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions", IF, Volume 105, 102218, ISSN 1566-2535, DOI:10.1016/j.inffus.2023.102218

[7] Brijesh Bakariya, et al., (2024), "Facial emotion recognition and music recommendation system using CNN-based deep learning techniques", ES 15, 641–658, DOI: 10.1007/s12530-023-09506-z

[8] Mustaqeem Khan, et al., (2024), "MSER: Multimodal speech emotion recognition using cross-attention with deep fusion", ESA, Volume 245, 122946, ISSN 0957-4174, DOI: 10.1016/j.eswa.2023.122946

[9] Parthiban Krishnamoorthy, et al., (2024), "A novel and secured email classification and emotion detection using hybrid deep neural network", IJCCE, Volume 5, Pages 44-57, ISSN 2666-3074, DOI: 10.1016/j.ijcce.2024.01.002

[10] Naveen Kumari, et al., (2023), "Saliency map and deep learning based efficient facial emotion recognition technique for facial images", MTA 83, 36841–36864, DOI: 10.1007/s11042-023-16220-0

[11] Hafiz Burhan UlHaq, et al., (2024), "Enhanced real-time facial expression recognition using deep learning", ATML, vol. 3, no. 1, pp. 24–35, DOI: 10.56578/ataiml030103

[12] Zhi Zhang, et al., (2024), "Torch EEGEMO: A deep learning toolbox towards EEG-based emotion recognition", ESA, Volume 249, Part B, 123550, ISSN 0957-417, DOI: 10.1016/j.eswa.2024.123550

[13] Sowmya B, et al., (2023), "Machine learning model for emotion detection and recognition using an enhanced Convolutional Neural Network", JIST, 12(4), 786, DOI: 10.62110/sciencein.jist. 2024.v12.786

[14] Trishita Dhara, et al., (2023), "A Fuzzy Ensemble-Based Deep learning Model for EEG-Based Emotion Recognition", CC 16, 1364–1378, DOI: 10.1007/s12559-023-10171-2

[15] Xianxun Zhu, et al., (2023), "Emotion recognition based on brain-like multimodal hierarchical perception", MTA 83, 56039–56057, DOI: 10.1007/s11042-023-17347-w

[16] Ketan Sarvakar, et al., (2023), "Facial emotion recognition using convolutional neural networks", MTP, Volume 80, Part 3, Pages 3560-3564, ISSN 2214-7853, DOI: 10.1016/j.matpr.2021.07.297.

[17] KishorBhangale, et al., (2023), "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network", Electronics 12, no. 4: 839, DOI: 10.3390/electronics12040839

[18] Daniel Yohanes, et al., (2023), "Emotion Detection in Textual Data using Deep Learning", PCS, Volume 227, Pages 464-473, ISSN 1877-0509, DOI: 10.1016/j.procs.2023.10.547

[19] AlaaAlslaity, et al., (2023), "Machine learning techniques for emotion detection and sentiment analysis: current state, challenges,

and future directions", BIT, 43(1), 139–164, DOI: 10.1080/0144929X.2022.2156387

[20] Premjeet Singh, et al., (2023), "Modulation spectral features for speech emotion recognition using deep neural networks", SC, Volume 146, Pages 53-69, ISSN 0167-6393, DOI: 10.1016/j.specom.2022.11.005

[21] https://www.kaggle.com/datasets/ananthu017/emotion-detection-fer Accessed on 12th June 2024

Table 5: Comparison Based on Various Optimization Algorithms

Model	*Optimization Algorithm*	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
VGG16	SGD	90.12	88.45	89.12	88.78
VGG19	RMSprop	91.35	89.6	90.45	90.02
ResNet50	Adam	89.98	87.89	88.76	88.32
ResNet101	SGD	92.1	90.34	91.22	90.78
InceptionV3	RMSprop	91.85	90.12	91	90.55
MobileNet	Adam	90.67	88.94	89.34	89.12
DenseNet	Adam	92.89	91.45	92.12	91.78
Xception	Adam	93.12	92.01	92.56	92.29
SEFEX (Proposed)	Adam	95.34	94.56	94.87	94.71
SEFEX (Proposed)	SGD	93.45	91.78	92.45	92.11
SEFEX (Proposed)	RMSprop	94.12	93.34	93.67	93.5

Table 6: Confusion Matrix for the Proposed SEFEX Model

Emotions	*Predicted Happiness*	*Predicted Sadness*	*Predicted Anger*	*Predicted Surprise*	*Predicted Fear*	*Predicted Disgust*	*Predicted Neutral*	*Predicted Contempt*
True Happiness	920	10	7	6	5	9	4	6
True Sadness	9	891	15	5	8	6	7	5
True Anger	7	13	900	8	10	7	6	5
True Surprise	6	5	4	935	9	5	4	5
True Fear	8	10	9	11	895	8	9	10
True Disgust	5	6	7	8	7	910	6	7
True Neutral	4	8	7	5	6	9	916	6
True Contempt	6	5	8	4	10	8	5	914