Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 06, June - 2025

Performance Evaluation of Facial Emotion Recognition (FER) Through Convolutional Neural Network Model

Srijeet Goswami,
Department of Electronics Engineerning,
Madras Institute of Technology,Anna University,
Chennai,TamilNadu,India.

Dr. Sridevi C, Associate Professor,
Department of Electronics Engineerning,
Madras Institute of Technology,Anna University,
Chennai,TamilNadu,India

Akash R,
Department of Electronics Engineerning,
Madras Institute of Technology,Anna University,
Chennai,TamilNadu,India.

Sudharsan K,
Department of Electronics Engineerning,
Madras Institute of Technology,Anna University,
Chennai,TamilNadu,India.

Abstract*— Facial Emotion Recognition (FER) has emerged as a crucial area in computer vision and artificial intelligence, enabling machines to interpret human emotions through facial expression analysis. This study focuses on developing and evaluating the performance of a Convolutional Neural Network (CNN) model for FER using the FER2013 and CK+ datasets. The FER2013 dataset comprises grayscale images captured in diverse real-world conditions with variations in lighting, occlusions, and facial orientations, making it a challenging benchmark. In contrast, the CK+ dataset consists of high-quality, posed facial expressions collected in controlled environments, providing a complementary perspective for performance evaluation. The proposed CNN model classifies fa- cial expressions into seven fundamental emotion categories: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. It integrates multiple convolutional layers for feature extraction, max-pooling layers for dimensionality reduction and overfitting prevention, and fully connected layers for high-level reasoning. A softmax output layer predicts emotion class probabilities. Advanced optimization techniques, including model checkpointing and early stopping, were employed to enhance training efficiency and generalization. Performance metrics such as accuracy, precision, recall, and F1-score were computed for a comprehensive evaluation of classification capabilities. Experimental results demonstrate the CNN model's effectiveness in FER, with comparative analyses conducted for both datasets. Findings reveal the impact of dataset characteristics—such as expression variability and environmental conditions—on model performance. This study underscores the importance of dataset diversity and model adaptability, providing valuable insights for future advancements in FER systems and real- world applications.

I. INTRODUCTION

Facial Emotion Recognition (FER) is an essential area of research in artificial intelligence and computer vision, aiming to equip machines with the ability to interpret human emotions through facial expressions. It has a broad range of applications, including mental health diagnostics, interactive learning environments, marketing analytics, and driver safety systems. By enabling automated emotion analysis, FER enhances human-computer interaction, making technological systems more intuitive and responsive to user needs.

The development of FER systems has been significantly advanced by deep learning techniques, particularly Convolutional Neural Networks (CNNs). CNNs have demonstrated remarkable success in image recognition tasks due to their ability to learn hierarchical patterns from image data. Unlike traditional feature extraction methods, CNNs can automatically capture spatial and temporal variations in facial expressions, improving the accuracy and efficiency of emotion recognition systems.

A major challenge in FER is the variability in facial expressions across different environments, demographics, and lighting conditions. To address this, researchers employ diverse datasets that encompass real-world scenarios and controlled laboratory settings. This study leverages two widely recognized datasets—FER2013 and CK+ (Cohn-Kanade Extended)—to evaluate the performance of a CNN-based FER model. FER2013 consists of a large-scale collection of facial images captured in natural settings, making it a robust benchmark for real-world applications. In contrast, the CK+ dataset contains high-resolution facial expression sequences recorded under controlled conditions, providing a valuable resource for analyzing subtle emotional transitions.

The primary objective of this study is to assess the efficiency and adaptability of a CNN-based FER model across different datasets by analyzing key performance metrics, including accuracy, precision, recall, and F1-score. The findings will contribute to improving the robustness and generalizability of FER systems, paving the way for their integration into various real-world applications, from healthcare and education to security and entertainment.

II. LITERATURE SURVEY

A. Custom Lightweight CNN Model for FER Mustafa Can Gursesli et al. proposed a Custom Lightweight CNN Model (CLCM) designed to be computationally efficient while maintaining high accuracy in emotion recognition tasks [1]. The model was evaluated on public datasets such as FER-2013, RAF-DB, AffectNet, and CK+, focusing on detecting seven emotional states. With only 2.3 million parameters, the model demonstrated competitive accuracy compared to MobileNetV2 (3.5M) and ShuffleNetV2 (3.9M), making it suitable for real-time and resource-constrained applications.

B. Real-Time Face and Emotion Recognition in Humanoid Robots Suci Dwijayanti et al. investigated the integration of face and emotion recognition in humanoid robots using CNN architectures such as AlexNet and VGG16 [2]. The dataset consisted of primary data collected from 30 electrical engineering students, resulting in 18,900 face recognition samples and 5,000 emotion recognition samples. VGG16 achieved higher accuracy (100% for face recognition and 73% for emotion recognition) compared to AlexNet (85% and 64%, respectively). The developed architecture showed promise for real-time robotic applications, with an error rate of only 2.52% in spatial positioning.

C. GA-SVM-Based FER Using Geometric Features Xiao Liu et al. introduced a FER system that leverages geometric features derived from facial landmarks, such as "landmark curvature" and "vectorized landmark" features [3]. The approach utilized a Genetic Algorithm (GA) for feature optimization and Support Vector Machine (SVM) for classification. The system achieved test accuracies of 95.85% (8-class CK+), 97.59% (7-class CK+), and 96.56%
(7-class MUG), outperforming CNN-based benchmarks in some cases. This study highlights the potential of combining traditional machine learning with feature selection for computationally efficient real-time applications.

D. Lightweight FER System for Visually Impaired Individuals Dina Shehada et al. proposed a lightweight FER system incorporating partial transfer learning to enhance adaptability and performance [4]. The approach utilized a custom CNN optimized for portability and efficiency, ensuring real-time usability. The integration of wireless connectivity further improved accessibility for visually impaired users, demonstrating the potential for practical applications in assistive technology.

E. FER in Educational Environments Using Machine Learning William Eduardo Villegas et al. explored the use of machine learning techniques to analyze facial gestures in teaching environments [5]. The developed system processes real-time visual data to assess students' emotional states, providing feedback to improve teaching methodologies. The study emphasizes the importance of AI-driven emotion recognition in enhancing personalized learning experiences.

F. Multimodal Emotion Recognition Jiahui Pan et al. introduced a multimodal emotion recognition system called Deep-Emotion, which integrates facial expressions, speech, and EEG signals [6]. The system employs an enhanced GhostNet for facial feature extraction, a lightweight fully convolutional neural network (LFCNN) for speech analysis, and a tree-like LSTM (tLSTM) for EEG processing. Decision-level fusion of these modalities improved recognition accuracy, outperforming unimodal approaches. The system demonstrated effectiveness on datasets such as CK+, EMO-DB, and MAHNOB-HCI, highlighting the growing importance of multimodal techniques in FER.

III. DATASETS USED

A. CK+ (Cohn-Kanade Plus)

The CK+ (Cohn-Kanade Plus) dataset is an extended version of the original Cohn-Kanade (CK) database and is a widely used benchmark for facial expression recognition. It contains 593 image sequences from 123 subjects, covering seven universal facial expressions: Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Contempt. Each image sequence begins with a neutral face and progresses to a peak expression, with detailed annotations of facial landmarks and action units (AUs). The CK+ dataset is designed for controlled experimental setups and is suitable for applications like micro-expression analysis and feature extraction. Its consistent resolution makes it ideal for training convolutional neural networks (CNNs).

B. FER2013

The FER2013 (Facial Expression Recognition 2013) dataset, introduced as part of the ICML 2013 challenge, is a large-scale dataset widely used for deep learning-based emotion recognition. It consists of 35,887 grayscale images, each of size 48x48 pixels, categorized into seven basic emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The dataset is divided into training (28,709 images), validation (3,589 images), and test (3,589 images) sets. FER2013 captures facial expressions in unconstrained environments, incorporating challenges like variations in lighting, occlusions, and facial orientations, making it suitable for real-world applications. Despite its diversity, the dataset poses challenges such as imbalanced class distribution, particularly with fewer samples for emotions like "Disgust." This makes it an excellent resource for evaluating the robustness of emotion recognition models in complex scenarios.

IV. METHODOLOGY

A. BLOCK DIAGRAM

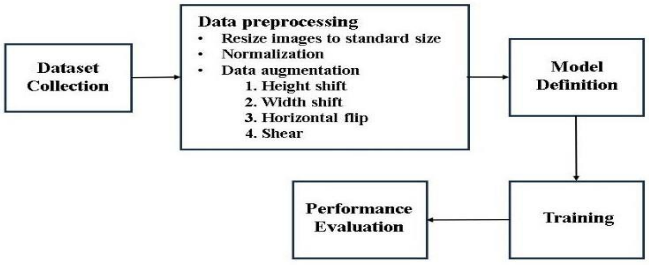


Fig. 1. BLOCK DIAGRAM

B. DATA PREPROCESSING

Data preprocessing is a crucial step in deep learning, as it prepares raw data to be suitable for training neural networks, enhances model performance, and ensures data consistency. One important aspect is data transformation, which includes normalization and standardization to scale features uniformly. Normalization scales values to a range, typically 0 to 1, while standardization adjusts data to have a mean of 0 and a standard deviation of 1. For categorical data, encoding methods like one-hot encoding (representing categories as binary vectors) or label encoding (assigning unique numerical values to categories) are applied to convert text labels into a numerical format suitable for machine learning models.

For image data, data augmentation helps reduce overfitting and improve generalization by artificially expanding the dataset through transformations such as rotation, flipping, scaling, cropping, and adjustments in brightness or contrast. Splitting the dataset into training, validation, and test sets is another critical step, ensuring the model is evaluated on unseen data to avoid overfitting. A typical split involves allocating 70 percent of data for training, 15 percent for validation, and 15 percent for testing.

Image data preprocessing also involves resizing images to a consistent shape, like 224x224x3 for models like VGG networks, and rescaling pixel values from their original range (0–255) to 0–1. This standardization speeds up training and prevents large pixel values from disrupting the learning process. In cases where the dataset is imbalanced, methods such as oversampling the minority class, undersampling the majority class, or assigning class weights during training can ensure better representation of all classes and improve the model's ability to handle underrepresented data. Lastly, shuffling data before training is important to eliminate any patterns in the order of data, preventing the model from overfitting to the sequence instead of the actual features.

Data augmentation is a technique used to artificially in- crease the size and diversity of a training dataset by applying various transformations to the existing data. It is especially valuable in deep learning for improving model generalization and reducing overfitting, particularly when the dataset is limited in size or lacks variability. By exposing the model to multiple variations of the same data, it learns to recognize patterns more robustly and becomes less sensitive to minor changes in input.

Common geometric transformations include rotation, translation, shearing, and flipping. These transformations modify the spatial layout of the image by rotating it by a specific angle, shifting it horizontally or vertically, or skewing it along an axis. Flipping the image horizontally helps simulate mirror images, which can be beneficial for object detection and recognition tasks.

In addition to geometric transformations, scaling and cropping techniques help diversify the training data by changing the size and focal point of the image. Zooming randomly alters the scale of the image, either by zooming in or out, while random cropping removes a portion of the image and resizes it back to the original dimensions. These methods simulate variations in object size and viewpoint, improving the model's ability to handle images at different scales.

Colour and intensity adjustments are another vital part of image augmentation, affecting the visual appearance of images to enhance model robustness. Brightness and contrast adjustments alter the light levels and visual sharpness, while saturation adjustments modify the intensity of colours, particularly in RGB images. Colour jittering combines all three adjustments—brightness, contrast, and saturation—in random proportions.

Other techniques, such as adding noise, applying affine transformations, and random erasing, simulate real-world imperfections by introducing distortions, random occlusions, or shifts in the image. These augmentations help the model generalize better in various environments and lighting conditions. Together, these preprocessing techniques create a robust pipeline for effective deep learning.

C. MODELS

Convolutional Neural Networks (CNNs): Convolu- tional Neural Networks (CNNs) are a class of deep learning models designed primarily for analyzing visual data. They are widely used for image classification, object detection, and other computer vision tasks. CNNs exploit spatial hierarchies in data using convolutional layers, pooling layers, and fully connected layers. CNNs consist of several key components designed to process and analyze image data effectively. The input layer accepts data, typically images represented as tensors, with dimensions indicating height, width, and channels (e.g., grayscale images with one channel or RGB images with three).

The convolutional layers apply filters (kernels) that slide over the input to detect features such as edges, textures, and patterns. These layers use parameters such as filter size (e.g., 3×3, 5×5), stride (step size for filter movement), and padding (zero-padding to preserve spatial dimensions). To introduce non-linearity, CNNs commonly use activation functions like ReLU (Rectified Linear Unit), which replaces negative values with zeros.

Pooling layers, following convolutional layers, down- sample feature maps and reduce spatial dimensions, allowing the network to focus on important features while reducing computational costs. Max pooling and average pooling are common types, where max pooling takes the maximum value from a window and average pooling computes the average. Dropout layers are also employed to prevent overfitting by randomly setting a fraction of the input units to zero during training, promoting generalization.

After convolutional and pooling layers, the data is flattened and passed through fully connected layers, where each neuron is connected to all others to facilitate the final classification. The output layer provides predictions using softmax activation for multi-class classification or sigmoid activation for binary classification. These components work together, enabling CNNs to extract features, learn representations, and make accurate predictions for tasks ranging from image recognition to object detection.

1) CNN Model 1: The CNN architecture for facial emotion recognition begins with an input layer that accommodates individual sample shapes. The architecture includes three convolutional layers: the first applies 64 filters with a 3×3 kernel and ReLU activation, while the second and

third layers use 32 filters each, maintaining the same kernel size and activation function. These layers capture hierarchical spatial features, such as edges and textures, which are vital for emotion classification.

Each convolutional layer is followed by a MaxPooling2D layer with a 2×2 pool size, reducing spatial dimensions and emphasizing prominent features, thus improving computational efficiency. The pooling outputs are then flattened into a 1D vector and passed to a Dense (fully connected) output layer with seven neurons. The softmax activation in the output layer generates probabilities for seven emotion classes (aligned with the CK+ dataset). The model is compiled using the sparse categorical crossentropy loss function for integer-labeled data and the Adam optimizer for efficient training. Accuracy is tracked as a performance metric, ensuring effective evaluation during training and testing.

2) CNN Model 2: The CNN model is created using the Sequential API in TensorFlow/Keras. This architecture consists of three convolutional layers, max-pooling layers, and a fully connected layer for classification. Key components include:

Convolutional Layers (Conv2D): These layers perform feature extraction by applying filters to the input image. The first layer uses 64 filters of size 3×3 with ReLU activation, and subsequent layers use 32 filters to refine and capture deeper features.

Pooling Layers (MaxPooling2D): These layers down- sample the feature maps by taking the maximum value within a specified window. This reduces the spatial dimensions and computational cost while preventing overfitting.

Flattening Layer (Flatten): The output of the final pooling layer is a multi-dimensional tensor, which is flattened into a 1D vector, preparing it for input into the fully connected layer.

Dense Layer (Dense): The fully connected layer maps the extracted features to the desired number of output categories, with softmax activation ensuring the output represents probabilities.

3) VGG19 Model: VGG19 is a deep convolutional neural network with 19 weight layers, comprising 16 convolutional layers and 3 fully connected layers. The architecture is simple and repetitive, making it easier to understand and implement. The VGG19 architecture consists of five convolutional blocks, each with multiple convolutional layers that use a 3×3 kernel and ReLU activation. The number of filters progressively increases, starting with 64 in Block 1 and scaling to 512 in Blocks 4 and 5. Each block concludes with a max-pooling layer (2×2 kernel, stride 2), which reduces spatial dimensions.

Block 1 starts with two convolutional layers (Conv1_1 and Conv1_2) with 64 filters, capturing basic patterns. Block 2 increases the filters to 128, followed by similar convolutional operations and max-pooling. Block 3, with four convolutional layers containing 256 filters, extracts more intricate features, while Blocks 4 and 5, each containing four convolutional layers and 512 filters, capture high-level representations and fine details.

The fully connected layers at the end of the network act as a classifier. The first two fully connected layers (FC1 and FC2) have 4096 neurons each with ReLU activation, enabling the network to learn complex combinations of features. The final layer (FC3) contains 1000 neurons with a softmax activation function, producing probabilities for each of the 1000 classes in the ImageNet dataset. VGG19's uniform design and depth make it highly effective for feature extraction and it is frequently used in transfer learning due to its pre-trained weights on large datasets.

D. Training Process

The training process of a deep learning model involves iteratively optimizing the model's parameters to minimize the error between its predictions and the actual outputs. It begins by passing the input data through the model in a forward pass, where the network computes predictions based on its current parameters. The error or loss is then calculated using a suitable loss function, such as categorical cross-entropy for classification tasks. Next, in the backward pass, the model uses backpropagation to compute gradients of the loss with respect to the model's parameters. These gradients are used by an optimization algorithm, like Adam or SGD, to update the parameters in a way that reduces the loss.

The process repeats for multiple epochs, where an epoch is one complete pass through the training data. To improve generalization and prevent overfitting, techniques like data augmentation and dropout are often applied. The model's performance is monitored on a separate validation set during training, allowing adjustments to hyperparameters or the use of early stopping if the validation loss stops improving. After training, the final model is evaluated on a test set to estimate its performance on unseen data. Proper visualization of metrics such as accuracy and loss throughout the process can help ensure effective and efficient training.

E. Performance Evaluation

The evaluation process of an image classification model involves assessing how well the model performs on a given task of classifying images into predefined categories. It is essential to determine how accurate and reliable the model is on both the training data (used for learning) and unseen test data (used for evaluation). Here's an overview of the key steps and metrics in the evaluation process:

1) Precision: Precision quantifies the accuracy of the model in predicting a specific emotion class. It is the ratio of true positive predictions to the total predicted positives for that class. High precision indicates a low rate of false positives.

$$Precision = \frac{TruePositives(TP)}{TruePositives(TP)+FalsePositives(FP)}$$

2) Recall: Recall measures the model's ability to correctly identify all instances of a particular class. It is the ratio of true positive predictions to the total actual positives for that class. High recall indicates a low rate of false negatives.

$$Recall = \frac{TruePositives(TP)}{TruePositives(TP)+FalseNegatives(FN)}$$

3) F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balanced metric when there is an uneven class distribution. It combines both precision and recall into a single measure, where a value close to 1 indicates excellent performance.

$$F1Score = 2 \ x \ \frac{Precision \times Recall}{Precision+Recall}$$

4) Support: Support refers to the number of actual instances of each class in the dataset. It provides insight into the class distribution, which is crucial when interpreting the model's performance.

V. RESULT AND DISCUSSION

The input requirements of various CNN models were analyzed, and preprocessing steps were tailored to optimize performance. Images from the datasets were converted to grayscale and resized to 224 x 224 pixels, ensuring compatibility with the models. Data augmentation techniques, including rotation, flipping, zooming, and brightness adjustments, were applied to expand the training data and reduce overfitting. This study evaluates the performance of different models on two datasets, CK+ and FER2013. Consistent preprocessing and augmentation were applied to both datasets for a fair comparison. The following sections provide a detailed analysis of the results obtained with each model on both datasets, highlighting their performance and limitations.

A. Classification Report

The classification report is a comprehensive summary that evaluates the performance of a classification model by providing key metrics such as precision, recall, F1-score, and support for each class. These metrics help to assess the effectiveness of the model in distinguishing between different emotion categories. The classification report is typically presented in a tabular format, where each row corresponds to an emotion class (e.g., happy, sad, angry) and columns represent the metrics. This report enables researchers to analyze the model's strengths and weaknesses for each class, identify areas for improvement, and ensure that the model performs consistently across all emotion categories. For example, in the context of facial emotion recognition using the CK+ and FER2013 datasets, the classification report can highlight how well the model distinguishes between subtle expressions (like fear and surprise) versus more distinct ones (like happiness and anger). It serves as a vital tool for model evaluation and optimization.

B. Confusion Matrix

A confusion matrix is a useful technique for evaluating the performance of a classification model, especially when dealing with multi-class datasets or datasets with imbalanced classes. It provides a summary of correct and incorrect predictions made by the model, breaking them down into count values for each class. Unlike classification accuracy alone, the confusion matrix offers deeper insights into the types of errors the model makes and helps identify specific areas for improvement.

The confusion matrix provides detailed information about the following: The number of true positive, true negative, false positive, and false negative predictions. Class-by-class performance, helping to assess the model's ability to differentiate between various emotions. The following parameters are used in a confusion matrix:

- True Positives (TP): Number of samples correctly predicted as belonging to a specific class.
- False Positives (FP): Number of samples wrongly predicted as belonging to a specific class when they do not.
- True Negatives (TN): Number of samples correctly predicted as not belonging to a specific class.
- False Negatives (FN): Number of samples wrongly predicted as not belonging to a specific class when they do.

VI. RESULTS

A. Cnn Model 1 Using Ck+ Dataset

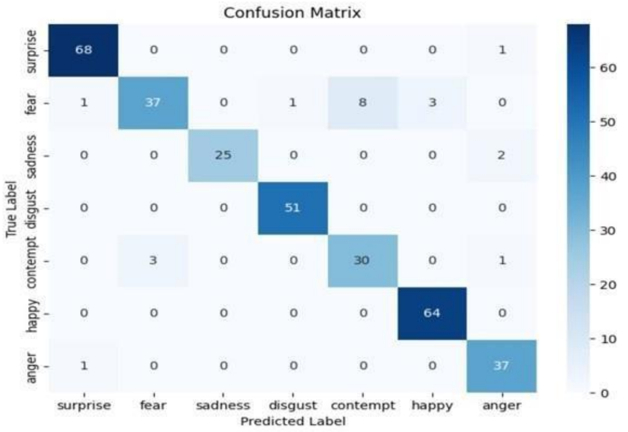
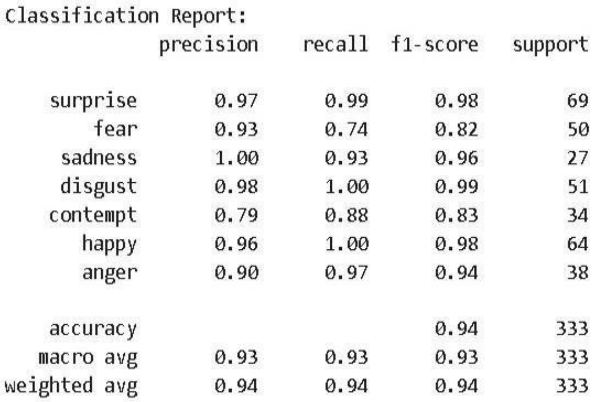


Fig. 2. CONFUSION MATRIX



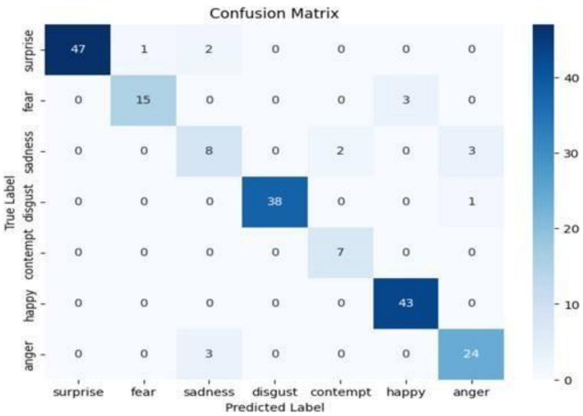
```
Classification Report:
              precision    recall  f1-score   support

    surprise       0.97      0.99      0.98        69
        fear       0.93      0.74      0.82        50
     sadness       1.00      0.93      0.96        27
     disgust       0.98      1.00      0.99        51
    contempt       0.79      0.88      0.83        34
       happy       0.96      1.00      0.98        64
       anger       0.90      0.97      0.94        38

    accuracy                           0.94       333
   macro avg       0.93      0.93      0.93       333
weighted avg       0.94      0.94      0.94       333
```

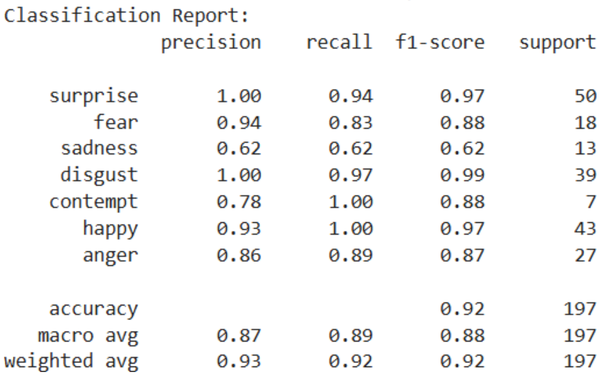
Fig. 3. CLASSIFICATION REPORT

B. Cnn Model 2 Using Ck+ Dataset



	surprise	fear	sadness	disgust	contempt	happy	anger
surprise	47	1	2	0	0	0	0
fear	0	15	0	0	0	3	0
sadness	0	0	8	0	2	0	3
disgust	0	0	0	38	0	0	1
contempt	0	0	0	0	7	0	0
happy	0	0	0	0	0	43	0
anger	0	0	3	0	0	0	24

Fig. 4. CONFUSION MATRIX



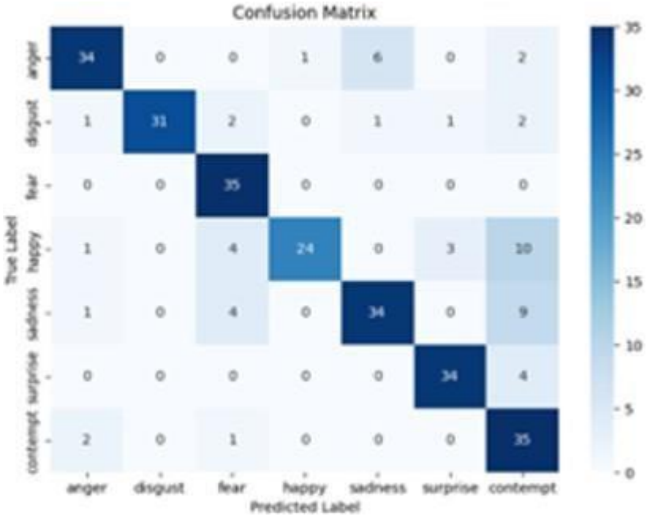
```
Classification Report:
              precision    recall  f1-score   support

    surprise       1.00      0.94      0.97        50
        fear       0.94      0.83      0.88        18
     sadness       0.62      0.62      0.62        13
     disgust       1.00      0.97      0.99        39
    contempt       0.78      1.00      0.88         7
       happy       0.93      1.00      0.97        43
       anger       0.86      0.89      0.87        27

    accuracy                           0.92       197
   macro avg       0.87      0.89      0.88       197
weighted avg       0.93      0.92      0.92       197
```

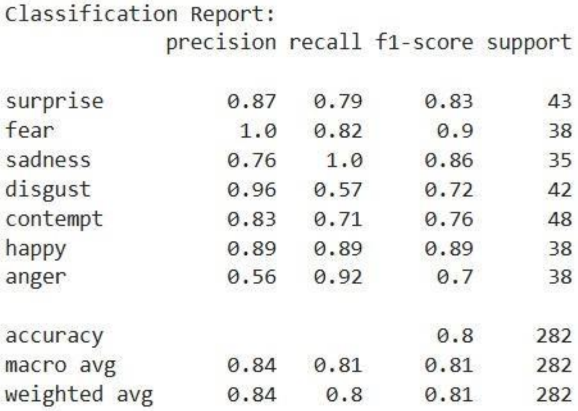
Fig. 5. CLASSIFICATION REPORT

C. Vgg19 Model Using Ck+ Dataset



	anger	disgust	fear	happy	sadness	surprise	contempt
anger	34	0	0	1	6	0	2
disgust	1	31	2	0	1	1	2
fear	0	0	35	0	0	0	0
happy	1	0	4	24	0	3	10
sadness	1	0	4	0	34	0	9
surprise	0	0	0	0	0	34	4
contempt	2	0	1	0	0	0	35

Fig. 6. CONFUSION MATRIX



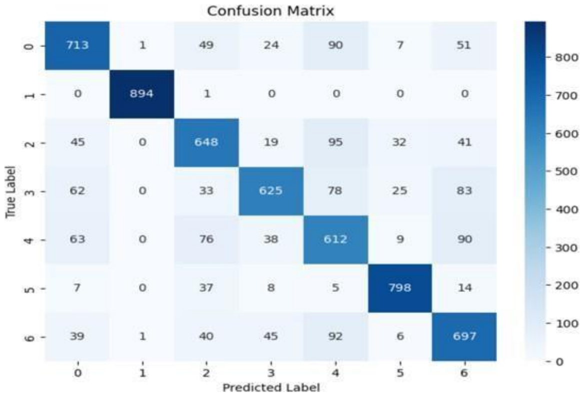
```
Classification Report:
              precision  recall  f1-score  support

surprise          0.87    0.79      0.83       43
fear               1.0    0.82       0.9       38
sadness           0.76     1.0      0.86       35
disgust           0.96    0.57      0.72       42
contempt          0.83    0.71      0.76       48
happy             0.89    0.89      0.89       38
anger             0.56    0.92       0.7       38

accuracy                             0.8      282
macro avg         0.84    0.81      0.81      282
weighted avg      0.84     0.8      0.81      282
```

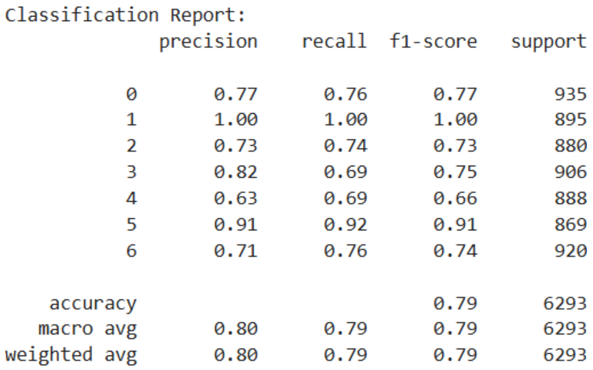
Fig. 7. CLASSIFICATION REPORT

D. Cnn Model 1 Using Fer-2013 Dataset



	0	1	2	3	4	5	6
0	713	1	49	24	90	7	51
1	0	894	1	0	0	0	0
2	45	0	648	19	95	32	41
3	62	0	33	625	78	25	83
4	63	0	76	38	612	9	90
5	7	0	37	8	5	798	14
6	39	1	40	45	92	6	697

Fig. 8. CONFUSION MATRIX



```
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.76      0.77       935
           1       1.00      1.00      1.00       895
           2       0.73      0.74      0.73       880
           3       0.82      0.69      0.75       906
           4       0.63      0.69      0.66       888
           5       0.91      0.92      0.91       869
           6       0.71      0.76      0.74       920

    accuracy                           0.79      6293
   macro avg       0.80      0.79      0.79      6293
weighted avg       0.80      0.79      0.79      6293
```

Fig. 9. CLASSIFICATION REPORT

E. Cnn Model 2 Using Fer-2013 Dataset

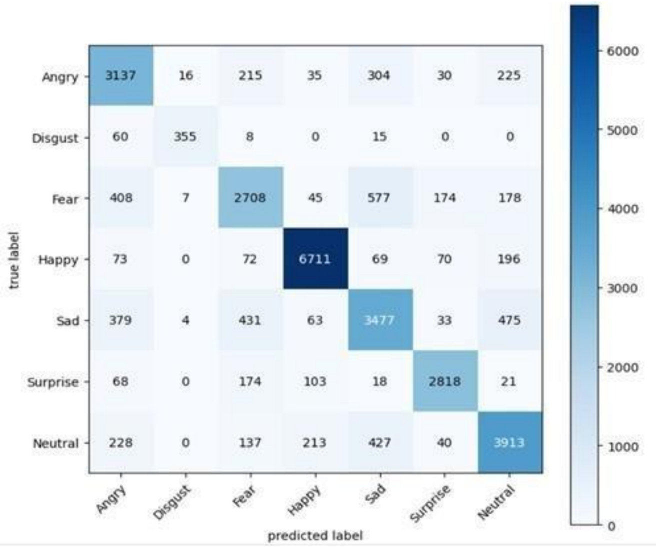
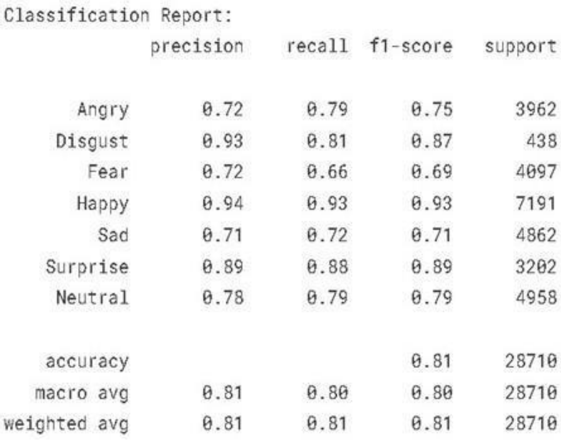


Fig. 10. CONFUSION MATRIX



```
Classification Report:
              precision    recall  f1-score   support

       Angry       0.72      0.79      0.75      3962
     Disgust       0.93      0.81      0.87       438
        Fear       0.72      0.66      0.69      4097
       Happy       0.94      0.93      0.93      7191
         Sad       0.71      0.72      0.71      4862
    Surprise       0.89      0.88      0.89      3202
     Neutral       0.78      0.79      0.79      4958

    accuracy                           0.81     28710
   macro avg       0.81      0.80      0.80     28710
weighted avg       0.81      0.81      0.81     28710
```

Fig. 11. CLASSIFICATION REPORT

F. Vgg19 Model Using Ck+ Dataset

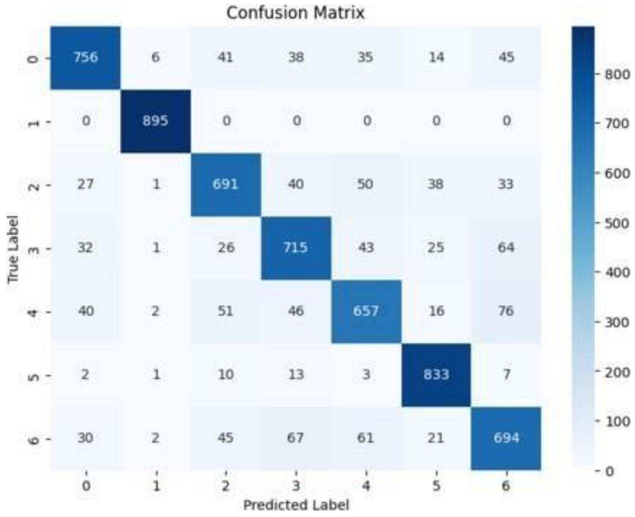
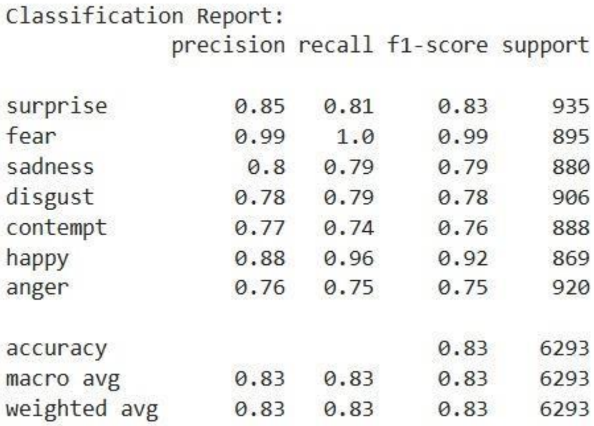


Fig. 12. CONFUSION MATRIX



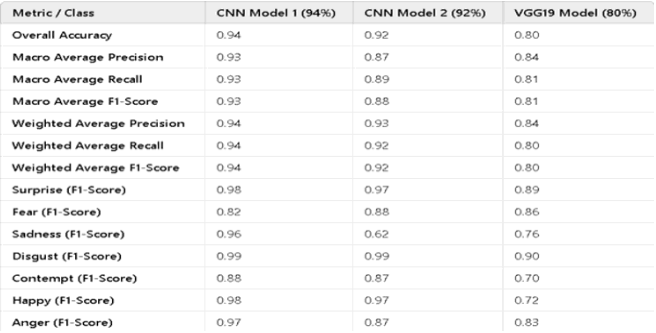
```
Classification Report:
              precision recall f1-score support

surprise           0.85   0.81     0.83     935
fear               0.99    1.0     0.99     895
sadness             0.8   0.79     0.79     880
disgust            0.78   0.79     0.78     906
contempt           0.77   0.74     0.76     888
happy              0.88   0.96     0.92     869
anger              0.76   0.75     0.75     920

accuracy                           0.83    6293
macro avg          0.83   0.83     0.83    6293
weighted avg       0.83   0.83     0.83    6293
```

Fig. 13. CLASSIFICATION REPORT

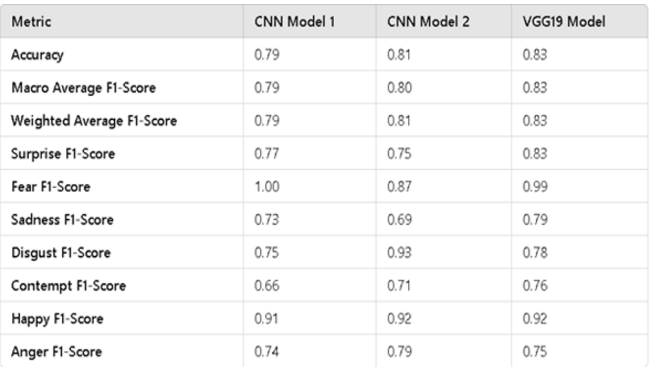
G. Comparison of Various Models for Ck+



Metric / Class	CNN Model 1 (94%)	CNN Model 2 (92%)	VGG19 Model (80%)
Overall Accuracy	0.94	0.92	0.80
Macro Average Precision	0.93	0.87	0.84
Macro Average Recall	0.93	0.89	0.81
Macro Average F1-Score	0.93	0.88	0.81
Weighted Average Precision	0.94	0.93	0.84
Weighted Average Recall	0.94	0.92	0.80
Weighted Average F1-Score	0.94	0.92	0.80
Surprise (F1-Score)	0.98	0.97	0.89
Fear (F1-Score)	0.82	0.88	0.86
Sadness (F1-Score)	0.96	0.62	0.76
Disgust (F1-Score)	0.99	0.99	0.90
Contempt (F1-Score)	0.88	0.87	0.70
Happy (F1-Score)	0.98	0.97	0.72
Anger (F1-Score)	0.97	0.87	0.83

Table. 1. COMPARISON OF VARIOUS MODELS FOR CK+

H. Comparison of Various Models for Fer2013



Metric	CNN Model 1	CNN Model 2	VGG19 Model
Accuracy	0.79	0.81	0.83
Macro Average F1-Score	0.79	0.80	0.83
Weighted Average F1-Score	0.79	0.81	0.83
Surprise F1-Score	0.77	0.75	0.83
Fear F1-Score	1.00	0.87	0.99
Sadness F1-Score	0.73	0.69	0.79
Disgust F1-Score	0.75	0.93	0.78
Contempt F1-Score	0.66	0.71	0.76
Happy F1-Score	0.91	0.92	0.92
Anger F1-Score	0.74	0.79	0.75

Table. 2. COMPARISON OF VARIOUS MODELS FOR FER-2013

VII. FUTURE SCOPE

The future of Facial Emotion Recognition (FER) us- ing Convolutional Neural Networks (CNNs) is evolving rapidly with advancements in deep learning, computing power, and new technologies. Real-time applications are a major focus, where CNNs are optimized for edge devices like mobile phones and IoT devices. These optimizations reduce model size and enhance speed, enabling FER on resource-constrained devices, including embedded systems and AR/VR devices, for real-time processing. Additionally, lightweight models are being developed to ensure efficient emotion detection without sacrificing accuracy in these environments.

Multimodal emotion recognition is a promising direction, combining FER with voice, physiological signals, and text analysis for more accurate emotion detection. This integration leverages techniques like transformers to fuse visual and non-visual cues, enhancing emotional context understanding. To improve real-world performance, CNNs must be able to handle challenges like lighting variations, occlusions (such as glasses or face masks), and facial orientation changes, which can affect recognition accuracy.

Additionally, models need to be culturally adaptive to recognize emotional expressions across diverse populations. Future advancements in FER also involve hybrid architectures, combining CNNs with Recurrent Neural Networks (RNNs) for better understanding of temporal and spatial emotion patterns. Attention-based models and Vision Transformers (ViTs) are being explored to focus on key facial features, while self-supervised learning is helping improve models with limited labeled data. FER has significant potential in Human-Computer Interaction (HCI), where it can personalize user experiences, such as tailoring virtual assistants or gaming systems based on emotional feedback. In therapeutic applications, FER can monitor emotional well-being and support mental health care. Data augmentation techniques like GANs and dynamic facial pose simulations are also enhancing the robustness and generalization of FER models, ensuring better performance in real-world scenarios.

VIII. REFERENCES

[1] Mustafa Can Gursesli , Sara Lombardi , Mirko Duradoni, Leonardo Bocchi ,Andrea Guazzini and Antonio Lanata ,"Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets"- IEEE Access -March 2024.

[2] Jiahui Pan, Weijie Fang, Zhihang Zhang, Bingzhi Chen, Zheng Zhang, Shuihua Wang, "Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG"- Sensors- Jan 2023

[3] Dina Shehada ,Ayad Turkey Wasiq Khan , Bilal Khan , and Abir Hussain, "A Lightweight Facial Emotion Recognition System Using Partial Transfer Learning for Visually Impaired People"- IEEE Access -April 2023.

[4] William Eduardo Villegas, Joselin García-Ortiz , Santiago Sánchez-Viteri, "Identification of Emotions From Facial Gestures in a Teaching Environment With the Use of Machine Learning Techniques"- Frontiers in Psychology, April 2023

[5] Z.-Y. Huang, C.-C. Chiang, J.-H. Chen, Y.-C. Chen, H.-L. Chung, Y.-P. Cai, and H.-C. Hsu, "A study on computer vision for facial emotion recognition", Sci. Rep., vol. 13, no. 1, p. 8425, May 2023.

[6] Puning Zhang, Miao Fu, Rongjian Zhao, Dapeng Wu, Hongbin Zhang, Zhigang Yang, and Ruyan Wang, "ECMER: Edge-Cloud Collaborative Personalized Multimodal Emotion Recognition Framework in the Internet of Vehicles"-IEEE Global Communications Conference (GLOBECOM) July 2023

[7] Suci Dwijayanti , Muhammad Iqbal, and Bhakti Yudho Suprapto, "RealTime Implementation of Face Recognition and Emotion Recognition in a Humanoid Robot Using a Convolutional Neural Network"-IEEE International Conference on Control, Automation, and Robotics (ICCAR),August 2022.

[8] M. Aziz and M. Aman, "Decision support system for selection of expertise using analytical hierarchy process method," IAIC Trans. Sustain. Digit. Innov., vol. 1, no. 1, pp. 49–65, Apr. 2021.

[9] Xiao Liu, Xiangyi Cheng, and Kiju Lee, "GA-SVM-Based Facial Emotion Recognition Using Facial Geometric Features"- Sensors - May 2021.

[10] S. Sharma and V. Kumar, "Performance evaluation of machine learning based face recognition techniques", Wireless Pers. Commun., vol. 118,no. 4, pp. 3403–3433, Jun. 2021.

[11] M. A. Hasnul, N. A. A. Aziz, S. Alelyani, M. Mohana, and A. Aziz, "Electrocardiogram-based emotion recognition systems and their applications in healthcare—A review", Sensors, vol. 21, no. 15, p. 5015, Jul.2021.

[12] R. Kobai and H. Murakami, "Effects of interactions between facial expressions and self-focused attention on emotion," PLoS ONE, vol.16, no. 12, Dec. 2021, Art. no. e0261666.

[13] M. N. A. Wahab, A. Nazir, A. T. Z. Ren, M. H. M. Noor, M. F. Akbar, and A. S. A. Mohamed, "Efficientnet-lite and hybrid CNN-KNN implementation for facial expression recognition on raspberry pi," IEEE Access, vol. 9, pp. 134065–134080, 2021.

[14] N. Agustini, N. Nursalam, T. Sukartini, G. K. Pranata, N. W. Suniyadewi, and I. D. A. Rismayanti, "Teaching methodologies regarding pallia- tive care competencies on undergraduate nursing students: A systematic review," J. Int. Dental Med. Res., vol. 14, no. 3, pp. 1302–1308, 2021.

[15] E. G. Krumhuber, D. Küster, S. Namba, D. Shah, and M. G. Calvo, "Emotion recognition from posed and spontaneous dynamic expressions: Human observers versus machine analysis," Emotion, vol. 21, no. 2, pp. 447–451, 2021.

[16] A. K. H. AlSaedi and A. H. H. Alasadi, "A new hand gestures recognition system," Indonesian J. Electr. Eng. Comput. Sci., vol. 18, no. 1, p. 49, Apr. 2020.

[17] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 18–31, Jan. 2019.

[18] S. Li and W. Deng, "Reliable crowdsourcing and deep localitypreserving learning for unconstrained facial expression recognition," IEEE Trans. Image Process, vol. 28, no. 1, pp. 356–370, Jan. 2019.

[19] K. R. Scherer, H. Ellgring, A. Dieckmann, M. Unfried, and M. Mor- tillaro, “Dynamic facial expression of emotion and observer inference,” Frontiers Psychol., vol. 10, p. 508, Mar. 2019.

[20] F. Makhmud Khujaev, M. Abdullah-Al-Wadud, M. T. B. Iqbal, B. Ryu, and O. Chae, “Facial expression recognition with local prominent directional pattern,” Signal Process., Image Commun., vol. 74, pp. 1–12, May 2019.

[21] J. Zhou, D. Yungbluth, C. N. Vong, A. Scaboo, and J. Zhou, “Estimation of the maturity date of soybean breeding lines using UAV-based multispectral imagery,” Remote Sens., vol. 11, no. 18, p. 2075, Sep. 2019.

[22] H.-Y. Lai, H.-Y. Ke, and Y.-C. Hsu, “Real-time hand gesture recognition system and application,” Sensors Mater., vol. 30, no. 4, pp. 869–884, 2018.

[23] S. Li, W. Deng, and J. Du, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2584-2593.

[24] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik, “The influences of emotion on learning and memory,” Frontiers Psychol.,vol. 8, pp. 1-22, Aug. 2017.

[25] B. Gaudelus, J. Virgile, S. Geliot, and N. Franck, “Improving facial emotion recognition in schizophrenia: A controlled study comparing specific and attentional focused cognitive remediation,” Frontiers Psychiatry, vol. 7, p. 105, Jun. 2016.

[26] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition” -2014

[27] I. J. Goodfellow et al., “Challenges in representation learning: A report on three machine learning contests,” in Proc. Int. Conf. Neural Inf. Process., in Lecture Notes in Computer Science, M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil, Eds. Berlin, Germany: Springer, 2013, pp. 117–124.

[28] R. Pandey and A. K. Choubey, “Emotion and health: An overview,” J. Projective Psychol. Mental Health, vol. 17, pp. 135–152, Jan. 2010.

[29] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2010, pp. 94101.

[30] L. K. McCorry, “Physiology of the autonomic nervous system,” Amer. J. Pharmaceutical Educ., vol. 71, no. 4, p. 78, Sep. 2007, doi: 10.5688/aj710478.

[31] J. M. Leppänen and C. A. Nelson, “The development and neural bases of facial emotion recognition”, in Advances in Child Development and Behaviour, vol. 34, R. V. Kail, Ed. San Diego CA, USA: JAI Publisher, Jan. 2006.