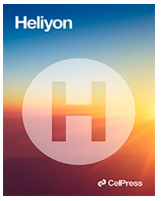
Heliyon 10 (2024) e38913

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Introducing a novel dataset for facial emotion recognition and demonstrating significant enhancements in deep learning performance through pre-processing techniques

Nursel Yalçin [a], Muthana Alisawi [b,c,*]

[a] Department of Computer and Instructional Technologies Education, Gazi Faculty of Education, Gazi University, Ankara, Türkiye
[b] Institute of Information, Computer Sciences, Gazi University, Ankara, Türkiye
[c] College of Education for Women, Kirkuk University, Kirkuk, Iraq

ARTICLE INFO

Keywords:
Deep learning architectures
Facial emotion recognition FER
FER13
Extended Cohn-Kanade (CK+)
Convolutional neural network CNN
Features extraction
Attention mechanisms

ABSTRACT

Facial expression recognition (FER) plays a pivotal role in various applications, ranging from human-computer interaction to psychoanalysis. To improve the accuracy of facial emotion recognition (FER) models, this study focuses on enhancing and augmenting FER datasets. It comprehensively analyzes the Facial Emotion Recognition dataset (FER13) to identify defects and correct misclassifications. The FER13 dataset represents a crucial resource for researchers developing Deep Learning (DL) models aimed at recognizing emotions based on facial features. Subsequently, this article develops a new facial dataset by expanding upon the original FER13 dataset. Similar to the FER + dataset, the expanded dataset incorporates a wider range of emotions while maintaining data accuracy. To further improve the dataset, it will be integrated with the extended Cohn-Kanade (CK+) dataset.

This paper investigates the application of modern DL models to enhance emotion recognition in human faces. By training a new dataset, the study demonstrates significant performance gains compared with its counterparts. Furthermore, the article examines recent advances in FER technology and identifies critical requirements for DL models to overcome the inherent challenges of this task effectively. The study explores several DL architectures for emotion recognition in facial image datasets, with a particular focus on convolutional neural networks (CNNs). Our findings indicate that complex architecture, such as EfficientNetB7, outperforms other DL architectures, achieving a test accuracy of 78.9 %. Notably, the model surpassed the EfficientNet-XGBoost model, especially when used with the new dataset. Our approach leverages EfficientNetB7 as a backbone to build a model capable of efficiently recognizing emotions from facial images. Our proposed model, EfficientNetB7-CNN, achieved a peak accuracy of 81 % on the test set despite facing challenges such as GPU memory limitations. This demonstrates the model's robustness in handling complex facial expressions. Furthermore, to enhance feature extraction and attention mechanisms, we propose a new hybrid model, CBAM-4CNN, which integrates the convolutional block attention module (CBAM) with a custom 4-layer CNN architecture. The results showed that the CBAM-4CNN model outperformed existing models, achieving higher accuracy, precision, and recall metrics across multiple emotion classes. The results highlight the critical role of comprehensive and diverse data in enhancing model performance for facial emotion recognition.

* Corresponding author. College of Education for Women, Kirkuk University, Kirkuk, Iraq.
E-mail addresses: nyalcin@gazi.edu.tr (N. Yalçin), myaseen.alisawi@gazi.edu.tr, muthanayaseen@uokirkuk.edu.iq (M. Alisawi).

https://doi.org/10.1016/j.heliyon.2024.e38913
Received 25 May 2024; Received in revised form 16 September 2024; Accepted 2 October 2024
Available online 4 October 2024

1. Introduction

The problem of recognizing human emotions based on facial features has long been of interest to psychologists, neuroscientists, and computer scientists, as it is the foundation of effective social interactions and communications. Recently, after the development of DL techniques, there was a breakthrough in the field of emotion recognition based on facial images [1]. DL techniques have shown remarkable performance in detecting emotions through facial features by simulating the structure and functioning of the human brain. Moreover, DL architectures can learn from complex patterns, extract features from large data sets, and generalize learning capabilities to new data [2]. Recognizing emotions through facial features using DL techniques has emerged as a promising research area. Its applications span across robotics, mental health applications, and human-computer interaction [3].

A trained observer can generally recognize facial expressions consistently and nearly instantly [4]. Conversely, On the other hand, the interpretation of such emotional expressions by automatic systems is generally complex and challenging and still has many unanswered issues which demand vast research-effort [5].

FER has applications in various disciplines, including medicine, social sciences, automotive and consumer electronics, human-machine interaction, and human-robot interaction [6]. Facial emotion recognition is a multi-step process. Initially, a face image is acquired through a live or recorded camera. Then, segmenting the face from the image. Subsequently, the detected face undergoes normalization to remove any distortions [7].

Traditional FER approaches heavily rely on content-based methods. These methods typically employ mathematical features, templates, or classifiers based on hand-crafted features and various learning methods. These methods manually extract facial features such as the eyes, nose, and mouth. These features are often employed in conjunction with supervised learning approaches, including support vector machines (SVMs) and decision trees (DTs). Additionally, Gabor wavelets and histograms are often used for feature extraction from facial images, which are then fed into classifiers. Although these methods are simple to implement, they typically achieve low accuracy [8,9].

Convolutional neural networks (CNNs) have made significant advancements in the field of FER tasks. However, these methods using CNNs don't really capture the complex and important features needed to tell the difference between different facial expressions from a wide angle. Therefore, there is still ample opportunity to enhance the performance of current CNN models for facial expression recognition (FER) [10].

In the early days, the psychological models were pioneering works in the recognition of FER, especially Paul Ekman's six basic emotions [11]. FER has continued to be an active area of research in psychology and related fields. Key research focuses have included facial action units (FACS) proposed by Ekman and Friesen, facial geometry kinetics (FGK), and the facial action coding system created by Cohen, Kanade, and Cohn. Additionally, other artificial intelligence techniques like fuzzy logic, hidden Markov models (HMMs), neural networks, and SVMs were utilized. These focus on making the gaps between categories bigger, as opposed to machine learning algorithms like Lazy K-Star [12], which look for similarities between cases. Other combinations of these techniques are also being used. These techniques have shown promising results in improving the accuracy and robustness of FER systems [8].

To address the limitations of the current FER methods, we propose a novel approach combining CBAM, 4-stage CNN, and EfficientNetB7-CNN. Then, we accurately evaluate these models using a novel dataset and pre-processing methods. Specifically, we examine contemporary DL methods that use facial features to identify emotions and investigate how our newly introduced dataset influences these methods. The core contributions of this work are threefold:

- Introduce the balanced FER2024_CK + dataset, which combines, preprocesses, and enhances existing datasets to improve performance and reliability.
- Evaluate various DL models on this dataset, including our proposed models (CBAM-4CNN and EfficientNetB7-CNN), highlighting their strengths and weaknesses.
- Provide a comprehensive cost-efficiency analysis of the proposed models, demonstrating significant enhancements in performance.

The article is structured as follows. Section 2 introduces some of the important related works. Our proposed methodology is detailed in Section 3.1, which is part of the border materials and methods section (Section 3). Experimental results are presented in Section 4 and followed by a detailed discussion in Section 5. Section 6 concludes the study and identifies our primary future direction.

2. Related works

Computer vision is a rapidly growing field that combines psychology, AI, and human-computer interaction. Recently, significant advancements have been made to develop systems capable of accurately recognizing human emotions facial expressions. This interdisciplinary field has practical implications for human-computer interaction, healthcare, and emotionally intelligent technologies.

Nawaf and Jasim [13] proposed a FER system using a CNN algorithm based on VGGNet. The model was trained on the FER2013 and FER + datasets, which were augmented to include additional images. The model validated its effectiveness in recognizing human emotion through a mean accuracy of 79 %. Also, the authors in Ref. [14] suggested an Emotion Recognition Convolutional Neural Networks (ERCNN) model designed specifically for identifying human emotions. Compared to the pre-trained models, ERCNN demonstrated its superiority in terms of accuracy, speed, and overall effectiveness. The ERCNN model achieved an accuracy of 87.133 % (82.648 %) in the public (private) test.

Punuri et al. [15] presented a new strategy derived from the Transfer Learning (TL) approach called EfficientNet-XGBoost. EfficientNet-XGBoost model integrates the strength of the EfficientNet and XGBoost algorithms. The authors demonstrated its superiority over the originality of the approach. In order to expedite the learning process of the network and address the issue of the vanishing gradient, they incorporated fully connected layers that utilize global average pooling, dropout, and dense operations. EfficientNet is optimized by substituting the higher dense layer(s) and integrating the XGBoost classifier, rendering it appropriate for FER. The suggested method has been thoroughly validated on four benchmark datasets: CK+, KDEF, JAFFE, and FER2013. To address the problem of data imbalance in certain datasets, including CK+ and FER2013, artificial data augmentation was employed using geometric modification techniques. Regardless of the characteristics of the datasets, the suggested strategy outperformed the counterparts, achieving accuracy rates of 100 %, 98 %, and 98 % for the first three datasets, respectively. However, the effectiveness of the proposed study is not as promising when trained and tested via FER2013 datasets (72.54%). Gupta et al. [16] introduced modified Inception-V3, VGG19, and RESENT50. To evaluate the models, the authors developed using three datasets: FER-2013, CK+, and RAF-DB. Proposed + ResNet-50 achieved the best performance of 73 %, 89 % and 76 %, respectively.

Choi and Lee [17] proposed a Deep Convolutional Neural Network (DCNN) ensemble classifier to enhance the recognition of facial expressions in uncontrolled environments. The approach employed a stochastic optimization technique to determine the weights of the ensemble, with the objective of minimizing energy and producing individual members. The DCNN ensemble classifier demonstrates competitive FER performance based on experiments conducted on three wild FER datasets (FER2013, SFEW2.0, and RAF-DB) and got an accuracy of 76.69 %, 58.68 %, and 87.13 %, respectively. Finally, Table 1 summarizes the related works mentioned in this section.

3. Materials and methods

3.1. Proposed method

As shown in Fig. 1, this section explains the proposed method, which comprises three main steps.

STEP 1 The first step is essential for preparing the dataset used for the purpose of training DL models. The first step consists of three stages: pre-processing, followed by image enhancement, and ending with the use of augmentation technology in order to obtain a balanced and improved dataset.

STEP 2 Here, a group of DL models will be applied to the new dataset, and based on the results obtained, the DL model that will achieve the highest accuracy will be chosen.

STEP 3 In the final step, optimization is performed on the chosen DL model to obtain higher accuracy and then recognize emotions according to the classification distributed into seven and ten categories.

3.1.1. EfficientnetB7_CNN model

To leverage (pre-trained) models accessible on the Kaggle platform, we propose a new architecture based on the EfficientNetB7 model. We initialize the base model with pre-trained weights from ImageNet and exclude the top classification layer to enable customization for our specific task. The proposed model architecture, EfficientNetB7_CNN, comprises multiple sequential layers, each with distinct functionalities (see Fig. 2), as outlined below:

- One layer of "*GlobalAveragePooling2D*": reduce spatial dimensions.

Table 1
Comparative review of related works.

Study	Year	Architecture	Used Dataset	Val_accuracy
[17]	2021	DCNN	FER2013, SFEW2.0, and RAF-DB	76.69 %, 58.68 %, and 87.13 %, respectively
[13]	2022	CNN based on VGGNet	FER2013, FER+	79 %
[18]	2022	3 stage CNN	FER2013	82 %
[14]	2023	ERCNN	FER2013, FER+	82.64 %
[15]	2023	EfficientNet-XGBoost	CK+ and FER2013	100, 72.5 % respectively
[10]	2023	FER-CHC	FER2013	74.68 %
[16]	2023	Proposed + ResNet-50	FER-2013, CK+, RAF-DB	73 %, 89 % and 76 %, respectively.
[19]	2023	SSF-ViT (L)	FER2013	74.95 %
[20]	2023	CNN-based Inception-v3 architecture	FER2013	73.09 %
[21]	2023	Xception Net	FER2013	77.92 %
[22]	2023	EmoNAS	FER2013	67.9 %
[23]	2023	SSA-NET	FER2013	67.57 %
[24]	2024	EduViT based on the MobileViT architecture	FER2013	66.51 %
[25]	2024	Hybridized CNN-LSTM	FER2013	79.34 %
[26]	2024	Activation-matrix Triplet loss and pseudo label with Complementary Information	FER2013	71.62 %
[27]	2024	EfficientNet	FER2013	58.41 %
[28]	2024	Custom CNN	FER2013	57.4 %

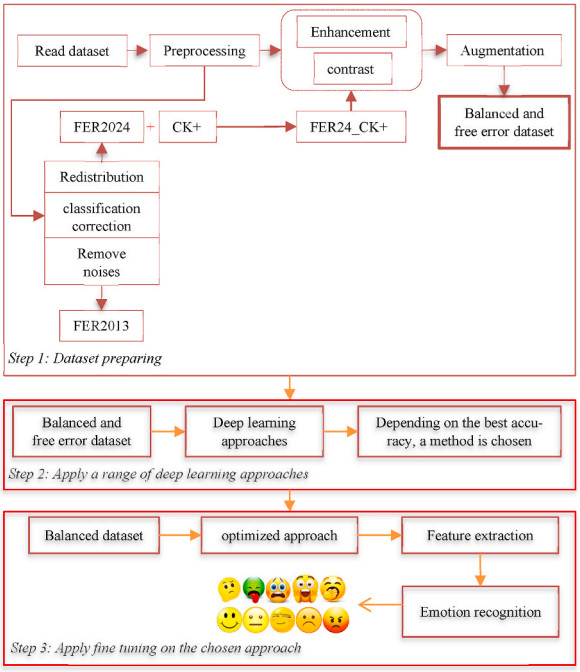


Fig. 1. The workflow of facial emotion recognition.

- Two layers of "*BatchNormalization*": stabilize training.
- A fully connected "*dense*" layer with 512 units and ReLU activation.
- The last layer is a "*SoftMax*" classifier tailored to the number of target classes.

Additionally, to optimize model training, we utilized the Adam optimizer with a learning rate of 0.0001 and the categorical entropy loss function to compile the model. The model was evaluated using accuracy, precision, and recall. To ensure that the model continues to be trained effectively, the "*ReduceLROnPlateau*" learning rate scheduling program was used. It calibrates the learning rate based on the validation loss, the amount of which is reduced by half if no improvement is observed within 15 epochs.

3.1.2. Convolutional block attention Module-4CNN (CBAM-4CNN) model

It can regard human attention as a tool that selects available processing resources, prioritizing task-relevant information in an input signal while attenuating irrelevant ones. CNNs have generalized such attention mechanisms to refine feature activations, demonstrating enormous potential in image recognition. A broad range of prior research has demonstrated that attention mechanisms offer enormous potential for advancing the performance of DCNNs. It can broadly categorize the attention mechanisms utilized in visual

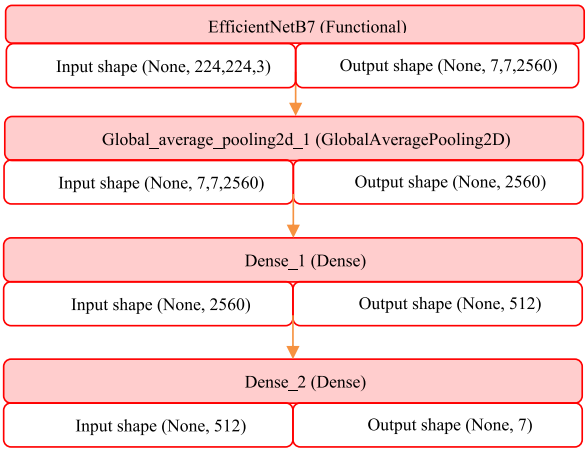


Fig. 2. EfficientnetB7_CNN model architecture.

recognition tasks into three groups: spatial attention, mixed-domain attention, and channel attention [29]. The Convolutional Block Attention Module (CBAM) is made up of channel attention and spatial attention modules. While the channel attention module emphasizes feature map channel weights, the spatial attention module focuses on the pixel regions within the image [30].

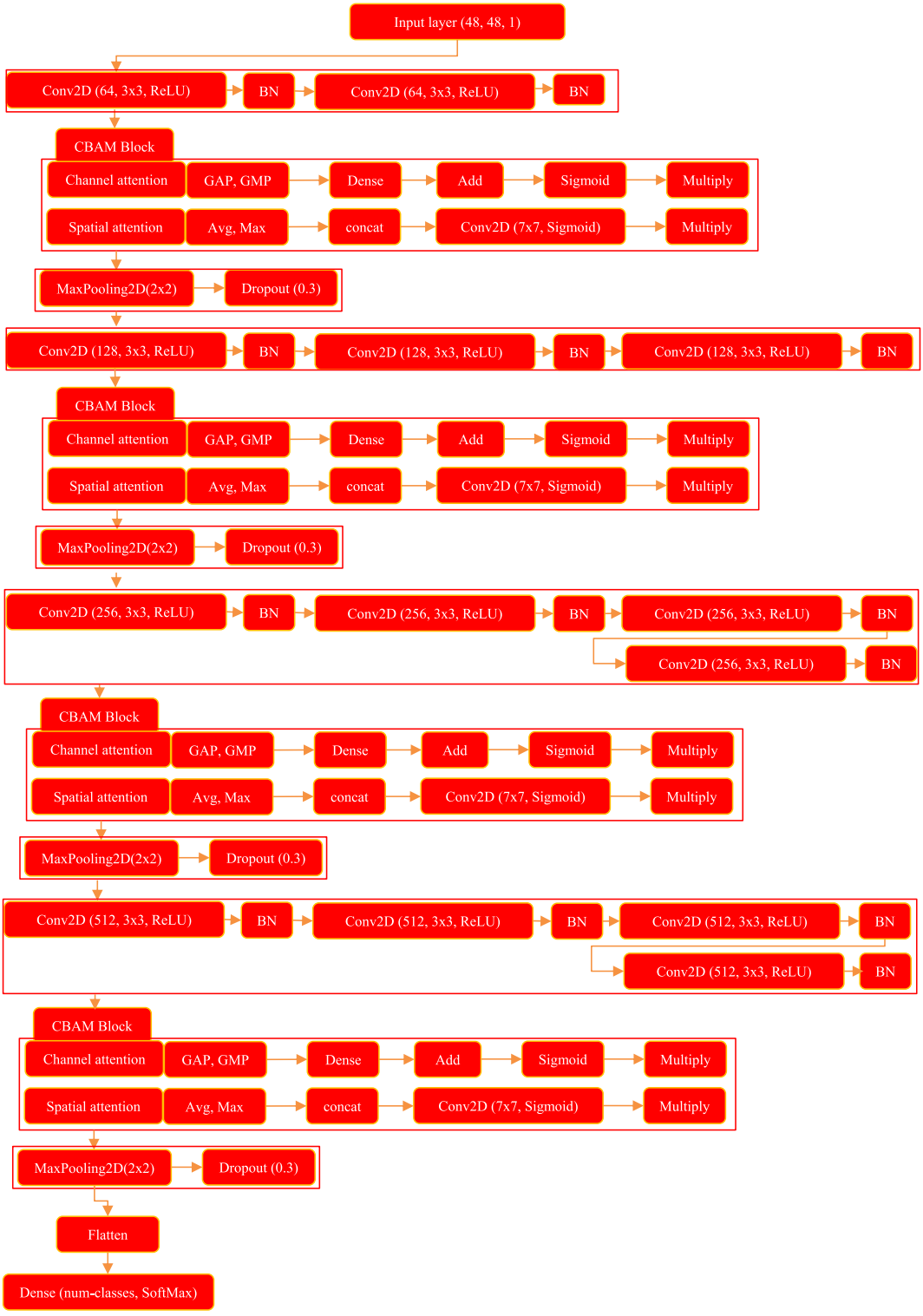


Fig. 3. CBAM_4CNN model architecture.

Therefore, to improve the representation of features for tasks that need to recognize facial emotions, we propose combining CBAM with a DCNN structure (CBAM-4CNN). The CBAM-4CNN model begins with an input layer tailored for 48 × 48 grayscale images. It uses several convolutional layers to pull out features, then batch normalization (BN) and CBAM blocks to make these features better by using channel and spatial attention. Each CBAM block leverages both average and max pooling to create a more focused feature map, enhancing the network's ability to capture salient details.

Our main goal is to improve recognition accuracy by effectively emphasizing relevant features while reducing the impact of noise and irrelevant information. Therefore, as shown in Fig. 3, the design has four convolutional stages, with dropout layers to stop overfitting and fully connected layers at the end to sort the extracted features into groups.

3.2. Datasets

To develop pre-trained and fine-tuned models, we utilized two datasets: CK + [31] and FER2013 [32]. These datasets are freely available for scientific research. Below, we give a short overview of the datasets.

FER13 [32]: The FER13 dataset is a widely used facial expression recognition dataset in the field of computer vision. It contains over 35,887 grayscale images (divided into 28709 for training and 7178 for testing) of faces labelled with one of seven different facial expressions. In the literature, the FER13 dataset has been used to train and evaluate different DL models for facial expression recognition.

After analyzing the FER13 dataset, the first observation is the diversity of expressions in the dataset. It includes both basic expressions (fear, happiness, anger, disgust, surprise, and sadness) and more subtle emotions (pride, embarrassment, and contempt). This diversity contributes to the development and training of models used to recognize emotions from facial features. In addition, it is important to note that there was a discrepancy in lighting, background, and face position in the image data set due to the images being taken in different conditions, such as camera angles and lighting settings in the background of the images. The accuracy of facial recognition algorithms can be negatively affected by these differences in facial image data. Therefore, careful consideration of these factors is crucial during the model training and evaluation process.

Moreover, the FER13 dataset classified its facial images based on the seven basic emotions, but incorrect classifications were observed for some image emotions. For example, images were classified as happy, but the correct classification was surprised.

From the preceding, the FER13 dataset is a valuable resource for researchers in the field of facial expression recognition. The diversity of expressions, lighting, and posture makes it a challenging yet realistic dataset for training and evaluating DL models. Nevertheless, variance in the dataset must be considered when developing and evaluating models.

On the other hand, a bias was observed in the dataset towards certain types of facial expressions, such as those that are more common or easier to recognize. This may lead to an imbalance in the distribution of facial expressions in the dataset, resulting in lower accuracy in recognizing fewer common expressions (Chart 1). Table 2 contains details of the FER-2013 dataset, the table shows the seven basic emotions used in this study.

The FER2013 dataset presented various obstacles, such as the inclusion of non-facial photos, inaccurate face cropping, partial occlusion, and inaccuracies in expression labelling. Finally, it is worth mentioning that these challenges have been extensively discussed in multiple articles (see Fig. 4).

Facial Expression Recognition Plus **(FER+):** The FER + annotations offer additional labels for the Emotion FER dataset. The previous figure displays the distribution of image numbers by emotion in the FER + dataset. The emotions of fear, contempt, and disgust are associated with a lower number of images, while the neutral feeling is associated with a larger number of images. The FER + dataset contains 35,710 photos, which is 177 images fewer than the original FER2013 dataset, which consists of 35,887 images. The discrepancy arises from the removal of the NF (Not Face) category and the exclusion of the unknown class, which comprises blurry photos (see Table 3).

FER + [33] dataset encapsulates a diverse spectrum of emotional states, including anger, sadness, fear, surprise, neutral, disgust, contempt, and happiness. The dataset serves as a valuable resource for training and evaluating facial expression recognition models. It offers researchers a robust resource to delve into the intricacies of emotion recognition technology. The inclusion of various emotional categories ensures the dataset's ability to address the multifaceted nature of human expressions. This makes it a pivotal asset for advancements in affective computing and computer vision research. The FER + dataset stands as a pivotal contribution to the field, fostering a deeper understanding of facial expressions and paving the way for the development of more nuanced and accurate emotion recognition systems. It is also noted that the bias and the imbalance in the categories of emotions in this dataset persists (Chart 2). Fig. 5 depicts examples of the FER13 vs. FER + labels.

Augmented CK + [31]: Includes a modified dataset consisting of 920 images derived from the original CK + dataset. The dataset has already been transformed into a 48 × 48 pixel size, with a grayscale color scheme, and cropped to include only the frontal face using the haarcascade_frontalface_default method [34]. The Haar classifier was utilized to enhance the visibility of images that were initially noisy due to variations in room lighting, hair shape, and skin color. These images represent eight distinct facial expressions:

Table 2
Distribution of emotional expressions in training and testing the FER13 dataset.

	Surprise	Fear	Angry	Neutral	Sad	Disgust	Happy	Total
Training	3171	4097	3995	4965	4830	436	7215	28709
Testing	831	1024	958	1233	1247	111	1774	7178

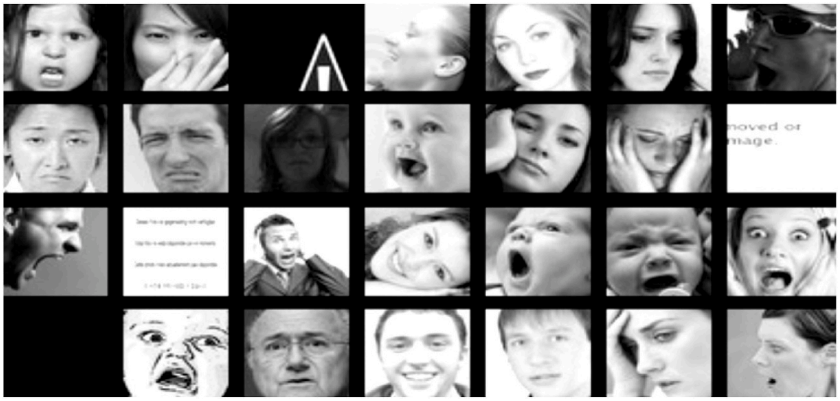


Fig. 4. Examples of some challenges in the FER2013 images.

Table 3
Distribution of emotional expressions in the FER + dataset.

Surprise	Fear	Angry	Neutral	Sad	Disgust	Happy	Contempt	Total
4493	825	3123	13014	4414	253	9367	221	35710

Fig. 5. FER + vs FER2013.

anger, contempt, disgust, fear, happy, sadness, neutral, and surprise (see Table 4). Chart 3 depicts the distribution of emotions in CK + classes.

3.3. FER2013 preprocessing

As reported in Ref. [35], the FER13 is a dataset from the wild, containing 35,887 images with a size of 48 × 48. It is considered one of the most widely used datasets in the field of facial emotion recognition. In this article, we focused on using this dataset after pre-processing it to eliminate all existing noise. As shown in Fig. 6, the pre-processing process was done manually by reviewing all 35, 887 images, as it was found that some images had been misclassified, for instance, 'Happy' expressions categorized as 'Angry' and so on. This process aims to obtain a dataset free of defects and errors in classification.

Through a comprehensive review of the FER13 dataset, many noises and defects were found in the established classifications of the classified emotions. Thus, the classification errors were corrected by redistributing the facial images to the correct classification that is appropriate for each image. In addition, all 94 images that did not represent facial images in both the training dataset and the investigation dataset were deleted. Moreover, three new categories were added that were not present in the previous data set (FER13),

Table 4
Distribution of emotional expressions in CK + dataset.

Surprise	Fear	Angry	Sad	Disgust	Happy	Contempt	Total
249	75	135	84	177	207	54	981

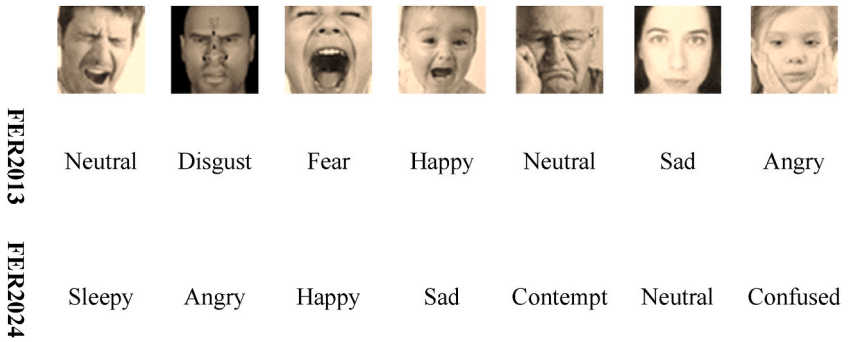


Fig. 6. Samples of errors in classification and its corrections in FER2024.

which are: contempt, confused, and sleepy (Tables 5 and 6).

3.4. New dataset FER2024

As a result of an accurate and comprehensive review of the original and wild datasets, FER13, a new dataset consisting of 10 categories (**angry, disgust, fear, happy, neutral, sad, surprise, contempt, confused, and sleepy**) with a total of 35,784 was found. Ninety-four images were deleted, which do not represent facial images (Table 7). Moreover, through the new and correct distribution of facial images, the difference in the total number of images for each category can be observed when compared with FER+. For example, but not limited to, the number of images of the emotion fear has become 3850 images (Table 7) instead of only 825 images in FER+ (Table 3).

3.5. Dataset combination

In order to enhance the new FER24 dataset with facial image samples that are free of defects and errors in classification, the CK + dataset was chosen. Thus, the number of facial images for each classification was increased (see Table 8).

The selected dataset (CK+) lacks classification for emotions recently added to the FER24 dataset, such as sleepy emotion.

As listed in Table 9; to begin using the new dataset, we redistribute it into two parts: 30 % (70 %) for testing (training).

Frequently, facial expressions misinterpret fear and surprise due to their similar key features, namely the eyebrows, eyes, and mouth. For example, when surprised, the eyebrows rise, exhibiting a greater curvature than when expressing fear. Additionally, when someone is surprised, their top eyelids and jaws tend to become more relaxed. Consequently, during this phase of the seven basic emotions, we used Ekman's concepts to classify emotions [36] (see Fig. 7). Particularly, the distinction between fear and surprise, to gain a more comprehensive understanding of emotions conveyed through facial images.

As a consequence, we obtained correctly classified and defect-free facial images at the end of the initial pre-processing (Tables 7 and 8). We have added two new classifications (unclear and eyeglasses) to the current classification. The first classification, under the 'unclear' label, contains a group of facial images that suffer from distortions that may negatively affect the process of training the model (such as blurring, cropped faces with incomplete features, and unclear emotions). The second classification, under the name Eyeglasses, contains a group of facial images with glasses that contain valuable information for recognizing certain expressions (as the glasses occlude the view of the eyes and eyebrows) [37], and sometimes even the region surrounding the eye. As a result, we isolated all 575 images (from testing 262 and training 313) that featured glasses (Fig. 8). Tables 10 and 11 show the results of the second pre-processing phase.

The final result produced 26,903 (dropping unclear, NHF, and eyeglasses images) flawless facial images. By merging these with the CK + dataset, which included 981 facial images, the total data expanded to 27,884 facial images. As listed in Table 12, to ensure exemplary training of the model, we re-divided these images into two main groups: 80 % (20 %) for training (testing).

Table 5
Correctly redistributed emotion samples for the FER13 dataset (training section).

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Contempt	Confused	Sleepy	Not Face
Angry	3625	15	27	20	55	119	22	46	52	0	14
Disgust	9	396	4	0	0	7	3	16	0	0	1
Fear	376	1	2729	51	267	115	60	42	447	0	9
Happy	5	0	5	7121	28	12	21	1	2	0	20
Neutral	12	1	14	206	4420	136	7	14	18	122	17
Sad	124	5	38	18	111	4479	6	28	11	0	7
Surprise	10	2	14	25	9	4	3097	1	0	0	9
New Total	**4161**	**420**	**2831**	**7441**	**4890**	**4872**	**3216**	**148**	**530**	**122**	**77**

Table 6
Correctly redistributed emotion samples for the FER13 dataset (testing section).

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Contempt	Confused	Sleepy	Not Face
Angry	956	0	0	0	0	0	0	0	0	0	2
Disgust	0	111	0	0	0	0	0	0	0	0	0
Fear	1	0	1019	0	0	0	0	0	2	0	2
Happy	0	0	0	1771	0	0	0	0	0	0	3
Neutral	0	0	0	20	1144	8	1	23	6	26	5
Sad	2	0	0	4	0	1238	0	1	0	0	2
Surprise	1	0	0	1	0	0	826	0	0	0	3
New Total	**960**	**111**	**1019**	**1796**	**1144**	**1246**	**827**	**24**	**8**	**26**	**17**

Table 7
Distribution of emotional expressions in the new FER2024 dataset.

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Contempt	Confused	Sleepy	Total
New distribution	5121	531	3850	9237	6034	6118	4043	172	538	148	**35784**

Table 8
FER2024&CK + combination.

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Contempt	Confused	Sleepy	Total
FER2024	5121	531	3850	9237	6034	6118	4043	172	538	148	35878
CK+	135	177	75	207	0	84	249	54	0	0	981
Total	5256	708	3925	9444	6034	6202	4292	226	538	148	36859

Table 9
FER2024&CK + new distribution.

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Contempt	Confused	Sleepy
Training 70 %	3680	496	2747	6612	4224	4342	3006	158	376	104
Private Testing 15 %	788	106	589	1416	905	930	643	34	81	22
Public Testing 15 %	788	106	589	1416	905	930	643	34	81	22

Fig. 7. Ekman's emotions expression concept [36].

Fig. 8. Samples of distorted facial images.

3.6. Deep learning architectures

Recently, DL architecture has become the most widely used within a short period in several fields. The primary factor contributing to its popularity is the exceptional accuracy of CNNs. Initially, DL provided the CNN architecture as a basic starting point for all subsequent architectures [38].

In recent years, GoogleNet, which includes ensemble learning, fully convolutional layers, and highly complex network architecture, has brought new ideas to the design of convolutional neural networks. Consequentially, many architectural improvisations improve computing efficiency by applying foundational concepts [39]. Furthermore, SqueezeNet uses a fully convolutional network to reduce the number of parameters and then applies the concept of CNN onset [40]. On the other hand, DenseNet implemented a revised approach that includes layer connections to address the problem of disappearing gradients in deep neural networks [41].

This section enumerates the key architectures in DL that were employed in this work on both the FER13 dataset and the newly modified FER24&CK + dataset. As mentioned earlier, transfer learning architectures refer to neural network designs that are specifically built or modified for transfer learning activities. Many neural network topologies are able to perform transfer learning accurately and guarantee the results that researchers expect.

However, due to their high efficiency in transferring knowledge across different tasks, some architectures are frequently used. Below are mostly utilized architectures [42]:

- **AlexNet** was an early example of a deep convolutional neural network architecture that successfully showcased the power of DL in the field of image classification. Pre-existing iterations of AlexNet that have been trained on extensive datasets, such as ImageNet, are frequently employed as tools to extract features for transfer learning tasks [43].
- The **Visual Geometry Group** (**VGG**) is a well-known convolutional neural network design that is recognized for its straightforwardness and efficiency. Transfer learning problems often utilize pre-trained iterations of VGG, specifically VGG16 and VGG19 [44,45].
- **ResNet**, short for Residual Network, is a neural network architecture that tackles the issue of the vanishing gradient problem in deep networks by including residual connections. ResNet50, ResNet101, and ResNet152 are very popular pre-trained variants of ResNet that are widely used for transfer learning due to their exceptional and accurate performance when dealing with image datasets [46].
- **Inception**, also referred to as GoogleNet introduced the inception module, which enables the network to capture features at various scales effectively. InceptionV3 and InceptionResNetV2, which are pre-trained versions of Inception, are commonly employed for transfer learning purposes [47].
- **MobileNet** is one of the technologies intended for developing vision applications that have limited processing resources, such as pre-trained MobileNetV2, which contributes to transfer learning in cases of scarcity of computational resources [48].
- The **Xception** model represents a modified version of the Inception transfer learning model, where depth-separable convolutional layers are added instead of regular convolutional layers. Xception pre-trained models are utilized for transfer learning applications, specifically when there is a need for both high accuracy and computational economy [49].
- **DenseNet** is a neural network architecture that establishes connections between every layer in a feed-forward manner. This design encourages the reuse of features and results in a more compact representation of the model. DenseNet121 and DenseNet169, which are pre-trained variants of DenseNet, are utilized for transfer learning tasks [50].
- **EfficientNet** is a neural network architecture that achieves a balance between model size and accuracy by scaling up the depth, breadth, and resolution of the network. EfficientNet models, ranging from EfficientNetB0 to EfficientNetB7, that have been pre-trained are becoming more commonly employed for transfer learning purposes [51].

4. Experimental results

In this work, we utilized the Kaggle environment for its facial image capabilities. To accelerate the computational tasks, we

Table 10
Correctly redistributed emotion samples for FER13 dataset (testing section) 2nd phase (NHF: not human face).

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Contempt	Confused	Sleepy	NHF	Unclear	Eyeglasses
Angry	2728	1	24	12	14	24	15	59	0	0	43	726	34
Disgust	0	319	0	0	0	0	0	0	0	0	0	0	0
Fear	52	0	1251	9	1	36	99	5	0	0	66	1202	27
Happy	5	0	1	6080	1	0	16	2	0	0	34	358	115
Neutral	5	0	0	94	2638	0	3	7	0	7	0	1412	58
Sad	51	0	9	13	60	2182	6	66	0	0	58	1877	18
Surprise	19	0	41	95	10	7	2262	0	0	0	30	485	57
Contempt	0	0	0	0	0	16	0	186	0	0	0	1	4
Confused	0	0	0	0	0	0	0	0	376	0	0	0	0
Sleepy	0	0	0	0	0	0	0	0	0	104	0	0	0
New Total	**2860**	**320**	**1326**	**6303**	**2724**	**2265**	**2401**	**325**	**376**	**111**	**231**	**6061**	**313**

Table 11
Correctly redistributed emotion samples for FER13 dataset (Testing section) 2nd phase.

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Contempt	Confused	Sleepy	NHF	Unclear	Eyeglasses
Angry	971	0	23	3	0	3	3	15	0	0	32	362	29
Disgust	20	120	12	2	0	2	0	12	0	0	10	25	9
Fear	87	0	361	21	4	30	35	4	0	0	25	523	13
Happy	1	0	0	2326	0	1	7	0	0	0	15	187	88
Neutral	3	0	0	73	1156	36	0	0	0	4	43	431	64
Sad	33	1	12	4	1	1373	4	9	0	1	56	252	30
Surprise	5	0	23	24	1	0	835	1	0	0	16	103	29
Contempt	0	0	0	0	0	9	0	15	0	0	0	0	0
Confused	0	0	0	0	0	0	0	0	162	0	0	0	0
Sleepy	0	0	0	0	0	0	0	0	0	44	0	0	0
New Total	**1120**	**121**	**431**	**2453**	**1162**	**1454**	**884**	**56**	**162**	**49**	**197**	**1883**	**262**

Table 12
FER2024&CK + new distribution 2nd phase.

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Contempt	Confused	Sleepy
Training 80 %	3292	494	1466	7141	3108	3043	2829	347	430	128
Private Testing 10 %	411	62	183	911	389	380	353	44	54	16
Public Testing 10 %	411	62	183	911	389	380	353	44	54	16
Total	**4114**	**618**	**1832**	**8963**	**3886**	**3803**	**3535**	**435**	**538**	**160**

employed a P100 GPU designed for image processing and neural networks. The Kaggle platform provided a storage space of 73 GB along with 29 GB RAM. Lastly, 16 GB GPU memory is utilized to handle the computations of the processing of facial images.

To ensure that all facial images were classified correctly, a confusion matrix was determined. As shown in Fig. 9, the confusion matrix and the obtained results revealed that the dataset still needs to be reclassified and redistributed. Additionally, we encountered data leakage during data analysis. Some images were incorrectly classified with multiple emotions, which negatively affected model accuracy (see Fig. 10).

4.1. Performance evaluation of different DL models

We utilized the FER2024&CK + dataset (26,903 samples) to determine the performance of fifteen models. Moreover, a set of parameters was applied uniformly for all methods used to indicate the extent to which defects in the facial image dataset affect the accuracy of each method, as detailed in Tables 13 and 14.

4.2. Augmentation and enhancement

In the enhancement process, we incorporated a variety of changes to improve the dataset. Using an alpha value of 1.5 and a beta value of 10, we achieved significant improvements in brightness adjustment, resulting in images with better illumination and visibility. Additionally, we optimized the contrast levels for better feature differentiation by applying a gamma correction with a gamma value of 1.5 and a gain of 1.0. We carefully calibrated these enhancements to balance brightness and contrast. This ultimately will contribute to improving model performance by providing clearer and more distinguishable images.

4.2.1. FER24_CK + augmentation

In the augmentation process, we incorporated a variety of changes to improve and balance the dataset (detailed in Table 12) by removing bias between classes (addressed in Chart 1). Using a rotation probability of 0.7 with a maximum left and right rotation of 10°, we achieved contrast in both scale and direction. Moreover, the horizontal flip method was used with a 0.5 probability to improve symmetry while maintaining adherence to the nearest adjacent filling position. To increase the range of spatial views, a random zoom operation with a 0.5 probability and a zoom percentage of 80 % was applied (Table 15). Finally, generating the specified number of augmented images helps in diversifying the dataset, ultimately improving the model performance.

As a result of the augmentation process, we obtained a balanced dataset according to the target number that we chose. It was 6000 samples in training and 1200 samples in testing for each class in the new dataset, as shown in Figs. 11 and 12, respectively. This technology provides a mechanism to overcome the bias present in the data set between the existing classes, which will provide more

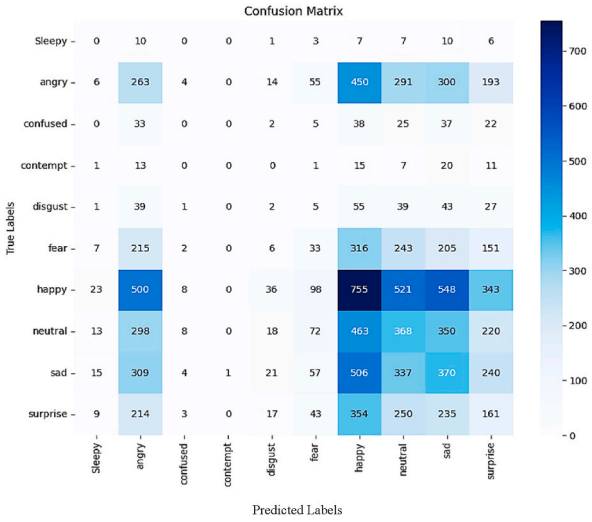


Fig. 9. Confusion matrix.

Fig. 10. Data leakage.

Table 13
Implemented key parameters in models' evaluation.

Parameter:	Optimizer	Learning rate	Drop rate	Loss function	Classifier
Value:	Adam	0.0001	0.5	Categorical_Crossentropy	SoftMax

Table 14
Evaluation of the models using FER13 and FER24_CK + datasets.

Model	Image size	Image type	Epoch	Batch size	Accuracy		
					FER13	FER24_CK+ 7 Emotions	FER24_CK+ 10 Emotions
VGG16	48 × 48	Grayscale	20	64	24.7 %	33.8 %	32.5 %
AlexNet	48 × 48	Grayscale	20	32	24.7 %	33.8 %	32.5 %
ResNet101	224 × 224	Grayscale	10	32	25.1 %	34 %	32.8 %
ResNet152	224 × 224	Grayscale	10	32	29.6 %	37.6 %	35.9 %
ResNet50	224 × 224	Grayscale	20	64	32.3 %	37.6 %	36.3 %
Standard CNN	48 × 48	Grayscale	50	64	37.4 %	45 %	42.2 %
InceptionV3	128 × 128	Grayscale	20	32	41.7 %	51.9 %	49.3 %
MobileNetV2	224 × 224	Grayscale	20	64	43.2 %	55.9 %	53.4 %
DenseNet121	48 × 48	Color	20	64	45.5 %	54.2 %	51.7 %
VGG19	48 × 48	Grayscale	20	64	56 %	67 %	62.6 %
Xception	71 × 71	Grayscale	20	32	62.6 %	69.9 %	69.5 %
EfficientNetB0	224 × 224	Grayscale	10	32	63.7 %	74.1 %	70.9 %
InceptionResNetV2	299 × 299	Grayscale	20	32	64.5 %	76.3 %	73.6 %
DenseNet169	224 × 224	Grayscale	20	32	65.4 %	78.4 %	74.2 %
EfficientNetB7	224 × 224	Grayscale	20	32	69.2 %	78.9 %	76.1 %

efficient training for the DL model.

In terms of the cost efficacy of the enhancement and augmentation processes, Tables 16 and 17 provide a detailed analysis of resource usage in two situations (10 and 7 emotions). It includes CPU and memory consumption, as well as the time spent for each stage of the dataset balancing process. This comprehensive analysis emphasizes the different computational requirements of different stages in the dataset preparation pipeline, highlighting the intensive nature of the balancing process compared to other stages.

4.3. Transfer learning EfficientNetB7-CNN

To accelerate the training, we leveraged the power of multiple GPUS on Kaggle. Additionally, to maximize the computational power available on Kaggle, particularly with GPU instances like T4 x2, it is crucial to distribute the training workload across GPUs. *TensorFlow's tf. distribute* provides an efficient way to achieve this. *MirroredStrategy* API. This strategy helps with synchronous training on multiple GPUs, where each GPU gets a copy of the model and processes a slice of the input batch. Consequently, using the MirroredStrategy, we can efficiently train DL models on multiple GPUs. Significantly increasing training speed and potentially improving model performance.

Transfer learning, a key component of DL techniques, lowers learning costs by leveraging knowledge from another task through training on a specific dataset [52]. As shown in Table 14, the transfer learning model EfficientNetB7 achieved the highest accuracy. As a result, we will work to develop this model through fine-tuning, as well as training it on the balanced and improved FER24_CK + dataset.

Table 15
FER24_CK+ (10 emotions) augmentation using Augmentor library.

Rotation	Value	Zoom_random	Value	Flip_left_right	Value
Probability	0.7	Probability	0.5	Probability	0.5
Max_left_rotation	10	Percentage_area	0.8		
Max_right_rotation	10				

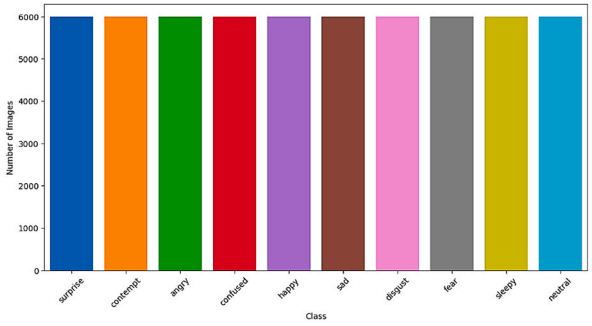


Fig. 11. Class distribution in the training set after balancing.

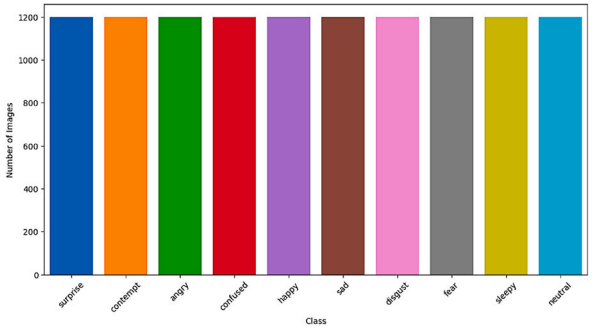


Fig. 12. Class distribution in the testing set after balancing.

Table 16
Resource usage and time analysis (10 emotions).

Stage	CPU Usage (%)	Memory Usage (%)	Time Taken (s)
Initial	0	4.2	
Visualization Before Balancing	0.5	4.2	2.96439
Sample Display Before Enhancement	0.7	4.2	5.48402
Dataset Augmenting (10 emotions)	15.3	4.5	296.95
Visualization After Balancing	0.7	4.5	298.707
Sample Display After Enhancement	0.7	4.5	301.132
Zipping Dataset	2	4.5	14.6526

Table 17
Resource usage and time analysis (7 emotions).

Stage	CPU Usage (%)	Memory Usage (%)	Time Taken (s)
Initial	0.5	4.3	
Visualization Before Balancing	0.5	4.3	2.92086
Sample Display Before Enhancement	3.5	4.3	5.21949
Dataset Augmenting (7 emotions)	9.8	4.4	164.799
Visualization After Balancing	0.5	4.4	166.406
Sample Display After Enhancement	2	4.4	168.699
Zipping Dataset	0.3	4.5	10.6167

We trained the EfficientNetB7 DL model on the Kaggle platform using 84000 samples in FER24_CK+ (10 emotions) and the T4x2 accelerator for 185 epochs. However, the training process stalled at 19/185 epochs. This accelerator provides two NVIDIA T4 GPUs, each providing 15 GB of GPU memory and significant computational power, designed to support end-to-end DL tasks. Despite these computational resources, numerous limitations impeded the training process's completion. In terms of insufficient GPU memory, EfficientNetB7 is a very complex multi-parameter model (64,426,398 params) that requires significant GPU memory for forward and backward propagation during training. Although allocating 15 GB of GPU memory per GPU, the complexity of handling high-dimensional data and model structure simultaneously necessitated additional memory resources. This limitation frequently resulted in out-of-memory (OOM) errors. Particularly, during the gradient accumulation phase, requiring the simultaneous storage of

activations and gradients. As for the limited continuous execution time, Kaggle provides 12 h per session, which is the maximum continuous execution time. Especially on large-scale datasets, training a state-of-the-art model like EfficientNetB7 requires long periods well beyond this threshold. This frequently interrupted the training process, requiring session restarts. These interruptions not only extended the overall training duration, but also presented challenges in maintaining consistency and progressing in training status.

To address this issue, we tried to use a transfer learning technique and implement freezing basis layers in EfficientNetB7. Nevertheless, we encountered a data mismatch issue during the pre-training phase of this model. This became apparent during the training phase, as the accuracy significantly decreased when we froze the layers in EfficientNetB7. In the first epoch, when using a dataset to identify emotions from facial images, the accuracy dropped from 64 % to 14 %. Since EfficientNetB7 was trained on a different dataset than the emotion recognition dataset (ImageNet) [53], the learned features from the frozen layers don't matter. Therefore, it makes the system less accurate. Therefore, we optimized the EfficientNetB7 DL model and trained it for only 20 epochs without freezing any layers.

The training introduced EfficientNetB7-CNN on the FER24_CK + dataset (7 classes) produced remarkable results over just 20 epochs, which took about 21600 s. The model initially performed modestly, as evidenced by its 32 % accuracy in the first epoch. However, as training progressed, both training and validation metrics showed significant improvement. As the epochs progressed in the training process, the accuracy of the model increased. The model demonstrated strong performance from the sixth epoch onward. The model achieved its peak accuracy of 93.75 % at the last twentieth epoch. This high accuracy demonstrates the model's ability to learn and generalize from the FER24_CK + dataset. Consequently, proving overall model performance as an accurate classification model. Table 18 displays the training process's implementation parameters.

Detailed analysis of the training results revealed that the model achieved a training accuracy of 93.66 %. On the validation set, the model achieved an accuracy of 78.72 %. The model's performance revealed variability across different categories (see Fig. 13 and Table 19). For example, the "Happy" category achieved precision, recall and F1-score 95 %, 92 % and 94 %, respectively. However, the Sad" group demonstrated lower performance metrics (64 %, 69 %, and 67 %, respectively). The overall accuracy for all categories combined was 79 %, with the averages of the overall and weighted accuracy, recall, and F1 scores of 79 %, 79 %, and 79 %, respectively.

4.4. CBAM-4CNN model

We chose a lightweight model (10,959,303 parameters) based on convolutional neural networks. It is improved by integrating it with an attention mechanism to overcome the problem of limited computational resources. We trained the model using a comprehensive dataset of 48,618 training images and two validation sets. Each set contained 4,200 images, all classified into seven distinct classes. The training process spanned 150 epochs, with a batch size of 50 and an initial learning rate of 0.0001. We used a sophisticated augmentation strategy to enhance the dataset's diversity and prevent overfitting. We used Adam's gradient descent optimizer and class cross-entropy as a loss function (see Table 20). We monitored key metrics like accuracy, loss, precision, and recall during training to assess the model's performance and steer potential hyperparameter adjustments.

The trained model achieved a maximum accuracy of 81.85 % during the training process. A detailed analysis of the training results revealed that the model achieved a training accuracy of 87.76 %, a precision of 90.28 %, and a recall of 85.25 %. On the validation set, the model achieved an accuracy of 77.48 %, a precision of 79.75 %, and a recall of 75.49 %. The model's performance revealed variability across different categories (see Fig. 14 and Table 21). For example, the "Happy" category achieved precision, recall and F1-score 93 %, 91 % and 92 %, respectively. However, the Sad" group demonstrated lower performance metrics (65 %, 60 %, and 62 %, respectively). The overall accuracy for all categories combined was 77 %, with the averages of the overall and weighted accuracy, recall, and F1 scores of 78 %, 77 %, and 77 %, respectively. The entire training process took about 12495 s.

According to the results, the proposed CBAM-4CNN model performs significantly better than many other state-of-the-art methods in this field. With an accuracy of 87.75 % and a validation accuracy of 77.48 %, the CBAM-4CNN model does better than most other methods. This shows how robust and useful it is for sentiment classification tasks. Overall, our CBAM-4CNN model has superior accuracy and validation performance over state-of-the-art methods, demonstrating its potential as a leading solution for sentiment classification tasks.

5. Discussion

Using multiple DL architectures on both the FER13 and modified FER24_CK + datasets yields intricate and subtle insights into emotion recognition (Table 15). Both VGG16 and AlexNet performed similarly in all instances, demonstrating that they have limited ability to adapt to the more complicated emotion classifications available in FER24_CK+. ResNet models showed modest improvements in accuracy as their topologies became more complex. ResNet50 beat other models in the FER13 and FER24_CK + datasets,

Table 18
EfficientNetB7-CNN (implementation parameters).

Input shape	Weights	Epochs	Batch size	Classifier	Optimizer	Loss function	Total used parameters
(224,224,3)	ImageNet	20	32	SoftMax	Adam	Categorical_Crossentropy	65,412,510

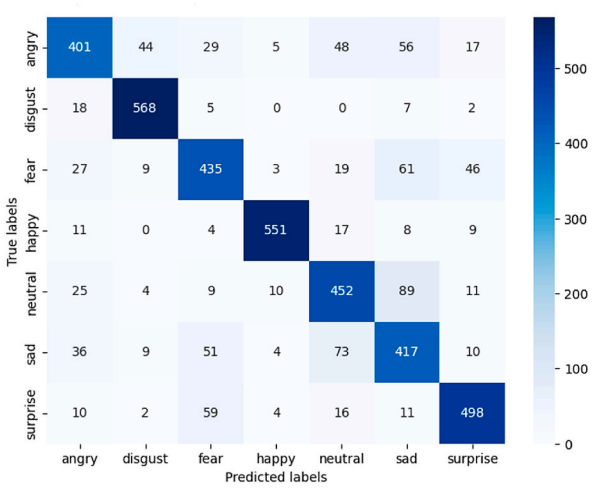


Fig. 13. Confusion matrix of EfficientNetB7-CNN for FER task on FER24-CK+ (7 classes) private testing.

Table 19
Outlines the EfficientNetB7-CNN performance measure for private testing.

Class	Precession (%)	F1-Score (%)	Recall (%)
Angry	76	71	67
Disgust	89	92	95
Fear	73	73	72
Happy	95	94	92
Neutral	72	74	75
Sad	64	67	69
Surprise	84	83	83

Table 20
CBAM-4CNN (implementation parameters).

Input shape	Epochs	Batch size	Classifier	Optimizer	Loss function	Dropout	Regularization	Total used parameters
(48,48,1)	150	50	SoftMax	Adam	Categorical_Crossentropy	0.25	Batch Normalization	10,959,303

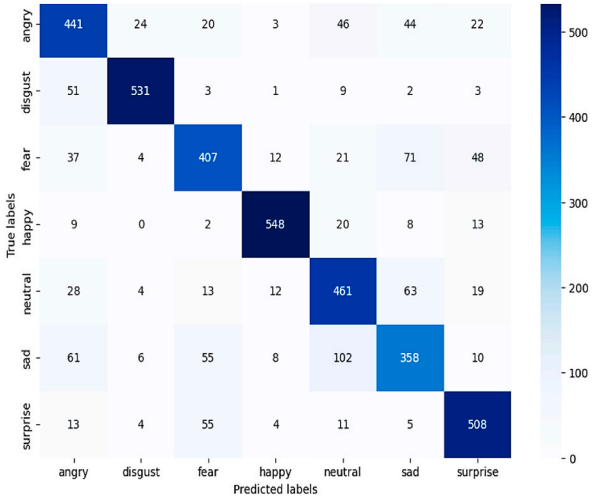


Fig. 14. Confusion matrix of CBAM-4CNN for FER task on FER24-CK+ (7 classes) private testing.

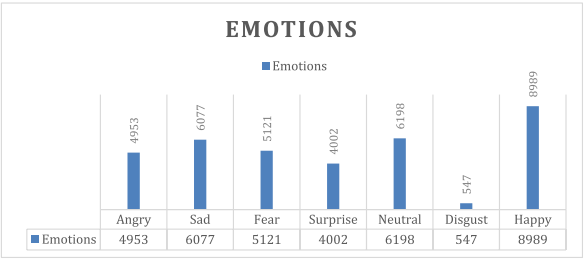


Chart 1. Bias in the distribution of classes samples in the FER13 dataset.

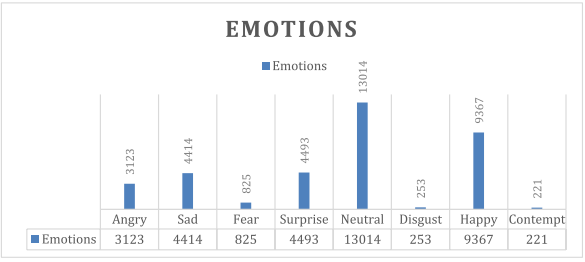


Chart 2. Bias in the distribution of classes samples in FER + dataset.

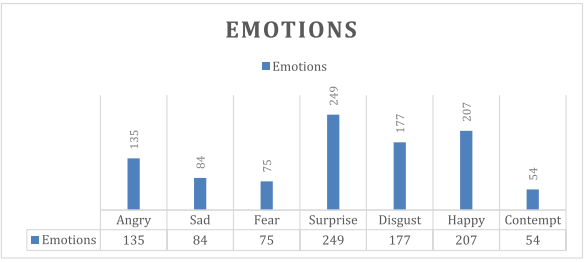


Chart 3. Bias in the distribution of classes samples in the CK + dataset.

which contain seven different emotions. Surprisingly, despite its simplicity, the CNN model has shown significant efficacy, particularly on the FER24_CK + datasets, demonstrating its capacity to tolerate dataset changes. However, the most notable advances were seen in advanced architectures like InceptionV3, MobileNetV2, and EfficientNetB0. These designs have shown great proficiency in capturing tiny emotional variations, as proven by their performance on the FER24_CK + dataset, which contains ten distinct emotions. The outstanding achievements of DenseNet169, EfficientNetB7, and InceptionResNetV2 demonstrate the need to use increasingly complicated architectures for comprehensive emotion recognition tasks. These findings emphasize the need to select appropriate DL models that are specifically adapted to the complexity of the dataset in order to attain peak performance in emotion recognition applications.

The results also showed that the EfficientNet models (B0 and B7) achieved better performance after being trained on the new FER24_CK + dataset. As listed in Table 19, the trained models outperformed the pre-trained models. These achievements were achieved without needing to improve either model's structure. Therefore, our results demonstrated the superiority of the EfficientNetB7 model to achieve more accurate results. EfficientnetB7_CNN achieved significant results within a span of just 20 epochs, constrained by the execution time and limited GPU resources available for this training process. Despite these limitations, the model demonstrated substantial improvements in precision, recall, accuracy, and loss. The training began with modest metrics, but through consistent epochs, the model reached a final accuracy of 93.75 %. Compared to other state-of-the-art approaches in Table 22, our proposed method, EfficientnetB7_CNN, achieves the highest recognition performance (78.72 % with only 20 training epochs), outperforming the most recent best performance [21] (i.e., 77 % with 60 training epochs). These promising results suggest that, with the removal of the current constraints on GPU resources and execution time, further training could potentially yield even superior performance. Extending the training period and leveraging more powerful computational resources would allow for deeper model refinement, likely enhancing its ability to generalize and perform effectively on unseen datasets.

The CBAM-4CNN model demonstrates reasonable levels of accuracy and robustness in recognizing different emotions, with a maximum accuracy of 81.85 %. As listed in Table 22, our proposed model clearly outperforms recent studies that rely on convolutional neural networks as the basis for the model, with an accuracy of 87.75 % and a validation accuracy of 77.48 %. Nevertheless, there is

Table 21
Outlines the CBAM-4CNN performance measure for private testing.

Class	Precession (%)	F1-Score (%)	Recall (%)
Angry	69	71	73
Disgust	93	91	89
Fear	73	70	68
Happy	93	92	91
Neutral	69	73	77
Sad	56	62	60
Surprise	82	83	85

notable variation in performance across different emotion categories. Categories such as "happiness" and "disgust" achieved high accuracy and recall scores, indicating that the model is particularly adept at recognizing these emotions. Conversely, the categories "sadness" and "fear" demonstrated lower accuracy and recall, indicating potential difficulties in accurately distinguishing these emotions (Table 21). This variation suggests that the model could benefit from further improvement and perhaps additional data or augmentation to improve performance in the lower-performing categories. Despite these challenges, the model demonstrates a solid foundation for emotion recognition, and targeted improvements could enhance its accuracy and reliability.

Comparing the FER13 dataset with the enhanced FER24_CK + dataset reveals notable differences in the performance of all designs. Therefore, the datasets significantly affect the accuracy. This confirms the crucial link between model design and dataset characteristics in emotion recognition tasks. This finding agrees with our previous study [35] that the success of the model depends critically on the composition of the dataset. Furthermore, the results listed in Table 14 demonstrate that the VGG16, AlexNet, and ResNet performed poorly on both datasets, with a slight improvement on the modified FER24_CK + dataset. Therefore, although these models can capture basic emotions well, their performance with the current architecture was not effective in both datasets compared to other networks used.

On the other hand, when moving to more advanced designs (such as InceptionV3, MobileNetV2, and DenseNet169), they have shown great efficiency in accuracy using the new FER24_CK + dataset. From this, we deduce the capability and adaptability of these architectures to extract intricate emotional characteristics from extensive datasets. Therefore, the quality of the dataset greatly affects the performance of DL models in emotion recognition tasks. Additionally, preprocessing and dataset enhancement led to a significant improvement in model performance. They also demonstrate the need to select and diversify data sets to improve model performance carefully.

6. Conclusion

This article delves into the field of facial emotion recognition through the lens of DL, with a primary focus on the FER13 dataset and the advances brought about by the new FER24 dataset. The article comprehensively provided an overview of the current landscape of facial emotion recognition. It studied the challenges and opportunities within the context while also presenting a depth analysis of recent evolution in DL models. By exploring several DL architectures on datasets such as FER13 and the enhanced FER24_CK+, it becomes clear that model complexity plays a critical role in achieving superior performance, especially on enriched datasets. The study

Table 22
State-of-the-art comparison of models' accuracy using the FER13 dataset as a base.

Backbone	Method name	Year	Accuracy	Val_accuracy
Pre-trained Models	**Ours (EfficientNetB7-CNN)**	–	93.66 %	78.72 %
	EfficientNet-XGBoost [15]	2023	Not reported	72.5 %
	Proposed + ResNet-50 [16]	2023	Not reported	73 %
	CNN-based Inception-v3 [20]	2023	Not reported	73.09 %
	Xception Net [21]	2023	Not reported	77.92 %
	FER-CHC [10]	2023	Not reported	74.68 %
	ResNet50 [27]	2024	59.41 %	54.67 %
	VGGNET [27]	2024	50.31 %	51.11 %
	EfficientNet [27]	2024	62.15 %	58.41 %
	SSER [26]	2024	Not reported	71.62 %
	ResNet50-CBAM-TCN [54]	2024	91 %	Not reported
	EduViT based MobileViT [24]	2024	Not reported	66.51 %
CNN based Models	**Ours (CBAM-4CNN)**	–	87.75 %	77.48 %
	Custom CNN [28]	2024	Not reported	57.4 %
	Hybridized CNN-LSTM [25]	2024	79.34 %	Not reported
	DCNN [27]	2024	82.56 %	65.68 %
	SSA-NET [23]	2023	Not reported	67.57 %
	EmoNAS [22]	2023	Not reported	67.9 %
	SSF-ViT(L) [19]	2023	74.95 %	71.7 %
	3 stage CNN [18]	2022	82 %	Not reported
	DCNN [17]	2021	Not reported	76.69 %

emphasizes the importance of dataset composition in determining model effectiveness and advocates the use of diverse datasets to enhance the accuracy of emotion recognition tasks. The creation of the FER2024 dataset (expanded emotion classes and integration with the CK + dataset) demonstrates the strides made in enhancing the dataset through fine-grained analysis and augmentation techniques.

Leveraging DL methodologies, especially CNNs, for emotion classification further contributes to the advancement of the field of facial emotion recognition. We believe that this work lays a solid foundation for future advancement in FER. In other words, by addressing dataset limitations, such as class imbalances and exploring the boundaries of DL applications, this article opens doors for further innovation in this exciting field. Furthermore, this study addressed the challenges of FER by introducing the FER24-CK + dataset. We used advanced preprocessing techniques to ensure data quality. By leveraging the EfficientNetB7 as a foundation and incorporating CNN optimizations, we developed a high-performance model capable of overcoming GPU memory constraints and achieving significant accuracy gains.

Additionally, we developed the CBAM-4CNN model, which integrates the convolutional block attention module with a custom 4-layer CNN architecture. CBAM-4CNN enhanced feature extraction and attention mechanisms. Our experimental results showed that the EfficientNetB7 model achieved a maximum accuracy of 93.75 %. CBAM-4CNN model outperformed with higher accuracy and recall across different emotion categories. Accordingly, our methods have shown significant progress in the field of emotion recognition, paving the way for more accurate and robust emotion recognition systems.

It is important to mention that the real-time applicability of the proposed model depends on several factors, including its accuracy, processing speed, and resource requirements [55]. The proposed models demonstrate promising accuracy across various emotion categories, making them promising models for real-time applications in fields requiring robust emotion recognition. Nevertheless, the variability in accuracy for some emotions, particularly the low performance in identifying “sadness,” may affect its real-time applicability in critical scenarios. Therefore, the model must be improved to ensure that it can generalize well to the diverse conditions that could be faced during real-time deployment. Regarding speed and processing, processing speed is a key factor for real-time applications. It is necessary to evaluate the inference time of the presented models to ensure their efficiency for real-time applications.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical approval

Not required.

Data and code availability

The original datasets utilized in this study are openly available. The implemented code in this article can be found on GitHub at the following link: https://github.com/FERProject/FER24_CKPlus/releases/tag/FER24_CK%2B. The repository contains the Balanced FER2024_CK + dataset that we have processed. Additionally, publicly available datasets, FER13 and CK+ were utilized. The following links were used to fetch the raw datasets: FER13 dataset: https://www.kaggle.com/datasets/msambare/fer2013; CK + dataset: https://www.kaggle.com/datasets/shuvoalok/ck-dataset.

CRediT authorship contribution statement

Nursel Yalçin: Writing – review & editing, Validation, Supervision. **Muthana Alisawi:** Writing – original draft, Methodology, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge assistant Prof. Ahmed Fakhrudeen (Kirkuk University), who assisted us with our article’s proofreading.

References

[1] A. Shabbir, M. Shabbir, M. Rizwan, F. Ahmad, Neuro-biological emotionally intelligent model for human inspired empathetic agents, J. Cogn. Syst. 4 (1) (2019) 1–11.
[2] M. Sari, A. Moussaoui, A. Hadid, Automated facial expression recognition using deep learning techniques: an overview, Int. J. Informatics Appl. Math. 3 (1) (2020) 39–53.
[3] E.S. Agung, A.P. Rifai, T. Wijayanto, Image - based facial emotion recognition using convolutional neural network on emognition dataset, Sci. Rep. (2024) 1–22, https://doi.org/10.1038/s41598-024-65276-x.

[4] G. Meena, K.K. Mohbey, A. Indian, M.Z. Khan, S. Kumar, Identifying emotions from facial expressions using a deep convolutional neural network-based approach, Multimed. Tool. Appl. 83 (6) (2024) 15711–15732, https://doi.org/10.1007/s11042-023-16174-3.
[5] J.Z. Wang, et al., Unlocking the emotional world of visual media: an overview of the science, research, and impact of understanding emotion, Proc. IEEE 111 (10) (2023) 1236–1286, https://doi.org/10.1109/JPROC.2023.3273517.
[6] C.M. Aqdus Ilyas, R. Nunes, K. Nasrollahi, M. Rehm, T.B. Moeslund, Deep emotion recognition through upper body movements and facial expression, Int. Jt. Conf. Comput. Vision, Imaging Comput. Graph. Theory Appl. 5 (January) (2021) 669–679, https://doi.org/10.5220/0010359506690679.
[7] S. Li, Application of entertainment e-learning mode based on genetic algorithm and facial emotion recognition in environmental art and design courses, Entertain. Comput. 52 (2024) 100798, https://doi.org/10.1016/j.entcom.2024.100798, 2025.
[8] F.Z. Canal, et al., A survey on facial emotion recognition techniques: a state-of-the-art literature review, Inf. Sci. 582 (2022) 593–617, https://doi.org/10.1016/j.ins.2021.10.005.
[9] N. Rodrigues, N. Costa, A. Pereira, Systematic review of emotion detection with computer vision and deep learning. Sensors, MDPI, 2024, pp. 1–29.
[10] X. Wu, J. He, Q. Huang, C. Huang, J. Zhu, X. Huang, FER-CHC : facial expression recognition with cross-hierarchy contrast, Appl. Soft Comput. 145 (2023) 110530, https://doi.org/10.1016/j.asoc.2023.110530.
[11] M. Heitkemper-Yates, A. Penjak, The practice of narrative: storytelling in a global context 3 (3) (2019).
[12] M.Y. Nawaf, M.M. Rashid, Study of data mining algorithms using a dataset from the size-effect on open source software defects, Kirkuk Univ. Journal-Scientific Stud. 15 (2) (2020) 25–44, https://doi.org/10.32894/kujss.2020.15.2.3.
[13] A.Y. Nawaf, W.M. Jasim, Human emotion identification based on features extracted using CNN, AIP Conf. Proc. 2400 (December) (2022), https://doi.org/10.1063/5.0112131.
[14] A.Y. Nawaf, W.M. Jasim, A pre-trained model vs dedicated convolution neural networks for emotion recognition, Int. J. Electr. Comput. Eng. 13 (1) (2023) 1123–1133, https://doi.org/10.11591/ijece.v13i1.pp1123-1133.
[15] S.B. Punuri, et al., Efficient net-XGBoost: an implementation for facial emotion recognition using transfer learning, Mathematics 11 (3) (2023) 1–24, https://doi.org/10.3390/math11030776.
[16] S. Gupta, P. Kumar, R.K. Tekchandani, Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models, Multimed. Tool. Appl. 82 (8) (2023) 11365–11394, https://doi.org/10.1007/s11042-022-13558-9.
[17] J.Y. Choi, B. Lee, Combining deep convolutional neural networks with stochastic ensemble weight optimization for facial expression recognition in the wild, IEEE Trans. Multimed. 25 (2023) 100–111, https://doi.org/10.1109/TMM.2021.3121547.
[18] D. Nixon, V. Vanjre, V. Petli, S. Hosgurmath, S.K. K, A Novel AI Therapy for Depression Counseling Using Face Emotion Techniques, vol. 3, 2022, pp. 190–194, https://doi.org/10.1016/j.gltp.2022.03.008. April.
[19] X. Chen, X. Zheng, K. Sun, W. Liu, Y. Zhang, Self-supervised vision transformer-based few-shot learning for facial expression recognition, Inf. Sci. 634 (March) (2023) 206–226, https://doi.org/10.1016/j.ins.2023.03.105.
[20] G. Meena, K. Kumar, S. Kumar, Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach, Int. J. Inf. Manag. Data Insights 3 (1) (2023) 100174, https://doi.org/10.1016/j.jjimei.2023.100174.
[21] K.K. Mohbey, G. Meena, K.K. Mohbey, Sentiment analysis on images using different transfer learning models, Procedia Comput. Sci. 218 (2023) 1640–1649, https://doi.org/10.1016/j.procs.2023.01.142.
[22] M. Verma, M. Mandal, S. Kumar, Y. Reddy, Efficient neural architecture search for emotion recognition, Expert Syst. Appl. 224 (March) (2023) 119957, https://doi.org/10.1016/j.eswa.2023.119957.
[23] Y. Liu, J. Peng, W. Dai, J. Zeng, S. Shan, Joint spatial and scale attention network for multi-view facial expression recognition, Pattern Recogn. 139 (2023) 109496, https://doi.org/10.1016/j.patcog.2023.109496.
[24] L. Quang, D. Trung, N. Chi, D. Thi, T. Thuy, Monitoring and improving student attention using deep learning and wireless sensor networks, Sensors Actuators A. Phys. 367 (2024) 115055, https://doi.org/10.1016/j.sna.2024.115055. October 2023.
[25] A.A. Bhat, S. Kavitha, S.M. Satapathy, J. Kavipriya, Real time bimodal emotion recognition using hybridized deep learning techniques, Procedia Comput. Sci. 235 (2024) 1772–1781, https://doi.org/10.1016/j.procs.2024.04.168.
[26] L. Pan, W. Shao, S. Xiong, Q. Lei, S. Huang, E. Beckman, SSER : semi-supervised emotion recognition based on triplet loss and pseudo label, Knowl. Base Syst. 292 (August 2023) (2024) 111595, https://doi.org/10.1016/j.knosys.2024.111595.
[27] D. Bhagat, A. Vakil, R. Kumar, A. Kumar, Facial emotion recognition (FER) using convolutional neural network (CNN), Procedia Comput. Sci. 235 (2023) (2024) 2079–2089, https://doi.org/10.1016/j.procs.2024.04.197.
[28] H. Manalu, A. Rifai, Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm, Intell. Syst. with Appl. 21 (2024) 1–18.
[29] Y. Yu, Y. Zhang, Z. Cheng, Z. Song, C. Tang, MCA : multidimensional collaborative attention in deep convolutional neural networks for image recognition, Eng. Appl. Artif. Intell. 126 (PC) (2023) 107079, https://doi.org/10.1016/j.engappai.2023.107079.
[30] F. Ma, Y. Li, M. Chen, Tactile texture recognition of multi-modal bionic finger based on multi-modal CBAM-CNN interpretable method, Displays 83 (October 2023) (2024) 102732, https://doi.org/10.1016/j.displa.2024.102732.
[31] Extended Cohn-Kanade (CK+)," [online]. Available: https://www.kaggle.com/datasets/shuvoalok/ck-dataset.
[32] Facial expression recognition 2013 dataset (FER2013)," [online]. Available: https://www.kaggle.com/datasets/msambare/fer2013.
[33] FERPlus (FER+)," [online]. Available: https://github.com/microsoft/FERPlus.
[34] J. Mcgrath, N. Nnamoko, TrackEd : an emotion tracking tool for e-meeting platforms, Softw. Impacts 17 (July) (2023) 100560, https://doi.org/10.1016/j.simpa.2023.100560.
[35] M. Alisawi, N. Yalçin, Real-time emotion recognition using deep learning methods: systematic review, Intell. Methods Eng. Sci. 2 (1) (2023) 5–21.
[36] P. Ekman, Universal Emotions | what Are Emotions? Paul Ekman Group, 2023. https://www.paulekman.com/universal-emotions/.
[37] D. Zeng, R. Veldhuis, L. Spreeuwers, A survey of face recognition techniques under occlusion, IET Biom. 10 (6) (2021) 581–606, https://doi.org/10.1049/bme2.12029.
[38] X. Wang, Y. Zhao, F. Pourpanah, Recent advances in deep learning, Int. J. Mach. Learn. Cybern. 11 (4) (2020) 747–750, https://doi.org/10.1007/s13042-020-01096-5.
[39] N. Yang, Z. Zhang, J. Yang, Z. Hong, J. Shi, A convolutional neural network of GoogLeNet applied in mineral prospectivity prediction based on multi-source geoinformation, Nat. Resour. Res. 30 (6) (2021) 3905–3923, https://doi.org/10.1007/s11053-021-09934-1.
[40] D. Apriadi, A.Y. Saputra, Modification of SqueezeNet for devices with limited computational resources, Resti (Rekayasa Sist. dan Teknol. Informasi) 7 (1) (2021) 19–25.
[41] A. Waheed, M. Goyal, D. Gupta, A. Khanna, A.E. Hassanien, H.M. Pandey, An optimized dense convolutional neural network model for disease recognition and classification in corn leaf, Comput. Electron. Agric. 175 (2020), https://doi.org/10.1016/j.compag.2020.105456.
[42] L. Alzubaidi, et al., Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, J. Big Data 8 (1) (2021), https://doi.org/10.1186/s40537-021-00444-8.
[43] W. Tang, J. Sun, S. Wang, Y. Zhang, Review of AlexNet for medical image classification, EAI Endorsed Trans. e-Learning 9 (2023) 1–13, https://doi.org/10.4108/eetel.4389.
[44] J. Cao, Artificial neural network models for image recognition, Highlights Sci. Eng. Technol. 62 (2023) 102–109, https://doi.org/10.54097/hset.v62i.10431.
[45] V. Sudha, T.R. Ganeshbabu, A convolutional neural network classifier VGG-19 architecture for lesion detection and grading in diabetic retinopathy based on deep learning, Comput. Mater. Contin. 66 (1) (2021) 827–842, https://doi.org/10.32604/cmc.2020.012008.
[46] M. Shafiq, Z. Gu, Deep residual learning for image recognition: a survey, Appl. Sci. 12 (18) (2022) 1–43, https://doi.org/10.3390/app12188972.
[47] L. Luo, P. Li, X. Yan, Deep learning-based building extraction from remote sensing images: a comprehensive review, Energies 14 (23) (2021) 1–25, https://doi.org/10.3390/en14237982.

[48] C.Y. Kim, K.S. Um, S.W. Heo, A novel MobileNet with selective depth multiplier to compromise complexity and accuracy, ETRI J. 45 (4) (2023) 666–677, https://doi.org/10.4218/etrij.2022-0103.
[49] S. Shirsath, V. Vikhe, P. Vikhe, Xception CNN-ensemble learning based facial emotion recognition, in: 2022 6th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2022, 2022, pp. 1–4, https://doi.org/10.1109/ICCUBEA54992.2022.10011111.
[50] L. Yin, P. Hong, G. Zheng, H. Chen, W. Deng, A novel image recognition method based on DenseNet and DPRN, Appl. Sci. 12 (9) (2022), https://doi.org/10.3390/app12094232.
[51] D. Masters, A. Labatie, Z. Eaton-Rosen, C. Luschi, "Making EfficientNet more efficient: exploring batch-independent normalization. Group Convolutions and Reduced Resolution Training, 2021.
[52] M. Iman, H.R. Arabnia, A Review of Deep Transfer Learning and Recent Advancements, Technologies 11 (2023) 1–14.
[53] M. Tan, R. Pang, Q. V Le, EfficientDet : scalable and efficient object detection, IEEE Xplore (2020) 10778–10787, https://doi.org/10.1109/CVPR42600.2020.01079.
[54] M. Aly, Revolutionizing Online Education: Advanced Facial Expression Recognition for Real-Time Student Progress Tracking via Deep Learning Model, 0123456789. Springer US, 2024.
[55] I.H. Sarker, Machine learning: algorithms, real-world applications and research directions, SN Comput. Sci. 2 (3) (2021) 1–21, https://doi.org/10.1007/s42979-021-00592-x.