Research Article

Facial Emotion Recognition and Classification Using the Convolutional Neural Network-10 (CNN-10)

Emmanuel Gbenga Dada,[1] **David Opeoluwa Oyewola**,[2] **Stephen Bassi Joseph**,[3] **Onyeka Emebo**,[4] **and Olugbenga Oluseun Oluwagbemi**[5]

[1]*Department of Mathematical Sciences, University of Maiduguri, Maiduguri, Nigeria*
[2]*Department of Mathematics and Statistics, Federal University Kashere, Gombe, Nigeria*
[3]*Department of Computer Engineering, University of Maiduguri, Maiduguri, Nigeria*
[4]*Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA*
[5]*Department of Computer Science, Faculty of Computer Science and Technology, Middlesex University, The Burroughs, NW4 4BT, London, UK*

Correspondence should be addressed to Emmanuel Gbenga Dada; gbengadada@unimaid.edu.ng

Received 20 March 2023; Revised 27 September 2023; Accepted 3 October 2023; Published 13 October 2023

Academic Editor: Mominul Ahsan

The importance of facial expressions in nonverbal communication is significant because they help better represent the inner emotions of individuals. Emotions can depict the state of health and internal wellbeing of individuals. Facial expression detection has been a hot research topic in the last couple of years. The motivation for applying the convolutional neural network-10 (CNN-10) model for facial expression recognition stems from its ability to detect spatial features, manage translation invariance, understand expressive feature representations, gather global context, and achieve scalability, adaptability, and interoperability with transfer learning methods. This model offers a powerful instrument for reliably detecting and comprehending facial expressions, supporting usage in recognition of emotions, interaction between humans and computers, cognitive computing, and other areas. Earlier studies have developed different deep learning architectures to offer solutions to the challenge of facial expression recognition. Many of these studies have good performance on datasets of images taken under controlled conditions, but they fall short on more difficult datasets with more image diversity and incomplete faces. This paper applied CNN-10 and ViT models for facial emotion classification. The performance of the proposed models was compared with that of VGG19 and INCEPTIONV3. The CNN-10 outperformed the other models on the CK + dataset with a 99.9% accuracy score, FER-2013 with an accuracy of 84.3%, and JAFFE with an accuracy of 95.4%.

1. Introduction

Facial expression is the apparent facial shift brought on by spontaneous reactions to emotional volatility [1]. Most of the time, it happens on its own and without warning. The automated facial expression entails using an artificial intelligence system to detect facial expressions in any situation [2, 3]. Facial recognition plays an important role in different applications [4–6]. Today, pattern recognition, computer vision, and its allied sciences have developed a significant attention in the field of facial expressions. Facial emotions may possibly depict the state of health or internal wellness of individuals. These facial emotions primarily fall into the categories of the seven classic ones: aggression, anxiety, astonishment, sadness, hatred, happiness, and indifference [7]. The process of determining human emotion is called emotion recognition [8]. The ability of people to accurately predict others' emotions differs considerably. Technology's application to helping people recognize emotions is a new field of research. The task of identifying a human's emotional experience is known as emotion detection [9]. There have been numerous decades of emphasis on human emotions.

Emotion studies began with Darwin's groundbreaking work on how humans and animals convey their emotions [10–12]. Multidisciplinary subjects like psychology, computer science, biology, sociology, and others have since made a substantial contribution. A major component of human interaction is facial expression, which instinctively conveys ideas and emotions to the other person. Since the face communicates a person's personality, feelings, sentiments, and beliefs before they are uttered, it plays a vital part in human social communication and interaction. The face is regarded as a critical component of the human being [13, 14]. This has resulted in the development of interactive systems with the goal of making the systems useable and convenient by concentrating on the users. This next generation of computing, termed captology, has produced enormous advantages that place the human user in the spotlight. The interfaces of this next generation computing quickly respond to human communication because they can recognize and comprehend the intentions and feelings that people express through social and affective signals [15]. These collaborative and interactive interfaces aim to modify person's attitudes and aspirations as well as enhance their health; as a result, persuasive spaces are created or persuasive technology is used to influence people's behavior or emotions in order to transform them into a preset state. This vision has generated enormous change in the fields of pattern recognition, computer vision, and human-computer interaction.

In social interaction, the face is crucial. Face biometrics are employed in a variety of business, investigative, and security applications. Identical to body language, face expressions are the quickest way to convey any kind of information. Happiness, sadness, rage, fear, disdain, and amazement are the six universally recognized basic expressions, according to a 1978 report by Ekman and Frisen. Automatic identification of facial expressions has found various applications in real-life situations, encompassing domains such as behavioural analysis, healthcare diagnosis, forensics (specifically falsehood identification), studies on commercial efficacy, and other related fields. The capacity of computers to analyze face expressions and therefore determine how people feel, as exemplified by smile detectors in commercialised digital imaging or interactive advertising, has introduced a novel dimension to interactions between humans and machines. Automatic identification of facial expressions holds promise in the context of robotic applications. The potential for robots to exhibit acceptable reactions and behaviours could be enhanced if they possessed the ability to anticipate how people feel.

However, there are a lot of issues that need to be solved before facial emotions in an image can be detected and recognized [16–22]. One issue is illumination, or how varied lighting settings can be seen in the input images depending on the strength of the camera used and the ambient lighting [23].

Operational difficulties with image recognition are widespread, as mentioned by other academics, and include challenges with system resources (memory), efficiency, reliability, and difficult solutions (nongeneric) [24, 25]. Therefore, the main goal of this work is to develop a system for recognizing and detecting facial emotions. Facial expression recognition can be quite difficult, and this is because of a number of things, including the 3D face posture, a lot of noise, opacity, and different lighting conditions. Face detection problems must be addressed since failing to detect the face or incorrectly detecting any of its parts would cause the system to malfunction, particularly for those systems that rely on or seek to extract features from regions of the face.

The scope of this work is limited to the application of CNN-10 and ViT models to human emotion recognition. For this work, only the CK+ (Cohn-Kanade+) dataset was employed, which is a widely used standard dataset in facial expression analysis and detection. The CK+ dataset comprises people's expressive reactions to various emotions shown in a controlled context. Because the dataset captures miniature variations in facial expressions, it is ideal for investigating sophisticated and delicate emotions. A classification mechanism that integrates gained knowledge about features into a label space that includes six designations: anger, contempt, disgust, fear, happiness, and sadness. Finally, the prediction model attempts to predict one of the six previously indicated labels for the input image. The proposed models are used in this procedure, and the model reduces an objective function that measures the difference between the prediction and the real label by finding the optimum mapping function.

The novelty of this work lies in the use of convolutional neural network-10 (CNN-10) and vision transformer (ViT) models for facial emotion detection. The work builds on previous advances in the discipline, including innovative aspects to improve performance and handle specific issues related to emotion recognition. The use of CNN-10 and ViT models for facial expression recognition is innovative since CNN-10 is a variation of CNN architecture intended specifically for facial expression recognition. Conventional CNN designs such as VGG and ResNet often incorporate additional layers, whereas CNN-10 is a simplified architecture that concentrates on obtaining critical features with fewer parameters. The uniqueness of the adopted approach lies in its ability to achieve superior performance while reducing its computational burden. Furthermore, CNN-10 includes a dynamic feature selection process in which the network learns to evaluate and pick biased facial features for expression detection. This adaptability improves the network's capacity to pay attention to crucial facial regions dependent on the circumstances surrounding various facial expressions, like the eyes or mouth. When contrasted with typical fixed-weight feature extractors, our technique provides better performance and enhanced feature exploitation.

The attention processes in ViT are another novel aspect of this work. Vision transformers (ViTs) are a revolutionary attention-based technique for recognizing facial expressions. ViTs record long-range dependencies between distinct facial regions via self-awareness processes, allowing for a more holistic comprehension of expressions. This attention technique allows the model to pay close attention to crucial regions while also taking global context into account, making it easier to recognize complicated expressions that

include interactions across numerous facial regions. Additionally, both the CNN-10 and ViT models increase face expression identification by combining local and global information. CNN-10 accomplishes this by successfully gathering local spatial data using convolutional filters, whereas ViTs capture global interactions between multiple facial patches with self-attention. The combination of local and global data allows for a greater understanding of facial expressions, which improves the models' discriminative power. Finally, the novelty of this work is that it can act as a precursor towards exploring and investigating the role that emotions may possibly depict in the mental health, the overall state of health, and the internal wellbeing of individuals.

This paper proposes deep learning for facial emotion recognition and detection to address these issues. The contributions of this work are outlined as follows:

(i) Developed a straightforward but effective method for facial emotion detection that uses deep convolutional neural networks and vision transformer (ViT) to detect faces from input images or videos

(ii) Application of CNN-10 and data augmentation for facial emotion recognition so that the model becomes more resilient and is less susceptible to small perturbations, thereby boosting its general robustness

(iii) Extraction of different facial features with detection accuracy of about 99.9% for CK+, 84.3% for FER-2013, and 95.4% for JAFFE datasets

(iv) Classification of facial expressions into diverse groups such as anger, contempt, disgust, fear, happiness, and sadness

(v) An evaluation of CNN-10 and VIT models' performance was done using important performance measures, including accuracy, confusion matrix, macroaverage and weighted-average, precision, recall, and $f1$-score. These measures were employed to assess the effectiveness of the models.

2. Related Works

Some works have been done on the area of facial emotion recognition. For example, Fan and Tjahjadi [26] proposed a technique for recognizing facial expressions that combines dynamic deep learning features with custom features. The suggested approach collects texture information from facial patches. The 593 segments in the CK+ dataset, which was created by 120 participants and contains seven basic facial emotions, were used by the authors. By storing shape, appearance, and deep dynamic information, the suggested model offers excellent efficacy and outperforms cutting-edge facial expression detection approaches on the CK+ dataset. The use of a comparatively small dataset by the proposed technique is a limitation of the work. Additionally, the system's performance was not evaluated against any well-known, highly effective deep learning method.

Minaee et al. [27] applied a deep learning approach that can concentrate on key facial features and outperform earlier models on a variety of datasets. Based on the results of the classifier, the authors also employed a modeling approach that may identify key facial regions for identifying various moods. The model was trained using 28,709 images, and the accuracy of the model was reported. On the test set, they were successful in achieving a very high accuracy rate. The limitation of the work is the inability of the proposed method to adequately cope with the imbalanced character of various emotion groups in the FER-2013 dataset.

Zhao et al. [28] proposed another facial expression recognition method. The strategy that the researchers suggested combines the DBNs' feature learning added benefit with the MLPs' classification benefit. The proposed model for facial expression recognition outperformed other cutting-edge classification techniques compared in the paper, as shown by the authors. The proposed models used the JAFFE database, which featured 10 Japanese women with seven different expressions. The limitation of the proposed work is that it cannot be used directly for classification and the effectiveness of their proposed method needs to be improved further still.

Abdulrahman and Eleyan [29] proposed support vector machines (SVMs) in facial expression recognition. Two different databases were used for the investigations. The results from all studies performed on these datasets show that the proposed approach has a satisfactory performance. The limitation of the work is that it finds it difficult to retain the key attributes with massive-scale configurations.

Rescigno et al. [30] proposed a personalized model through transfer learning for recognizing facial emotions. To extract the emotions of image features, the authors developed subject-specific models; the paper suggests using transfer learning. Pretty excellent performances were attained for the valence and arousal dimensions (RMSE for valence and arousal, respectively, was 0.09 and 0.1). General findings revealed that even though they alternated in making the predominant contribution, the authors made use of the AffectNet database, which houses more than 1 million samples obtained by doing searches on the Internet using keywords associated with emotions. The limitation of the work is that the pretrained net performed quite badly, indicating that accurate arousal level recognition depends more on the unique subject and is hence harder to extrapolate.

Jain et al. [31] proposed another deep learning-based approach for recognizing facial emotions. The proposed model was used on a network that has deep residual blocks and convolutional layers. In the proposed approach, all faces have had their image labels assigned for training purposes before the proposed DNN model is applied to the images. There are 8363 total images in the experimental dataset. Convolution layers in the suggested model extract features in a systematic order and are connected directly, and SoftMax layers are employed to denote six expression classes. The combined results demonstrate that the proposed model can perform better than the current front-line methods for emotion recognition. The weakness of the work is that the accuracy is still low and needs further improvement.

Akhand et al. [32] described an approach for recognizing emotions that leverages transfer learning in deep CNN. The proposed approach is validated using eight already-trained DCNN models and two popular face image datasets. With already-trained models, the proposed method demonstrated outstanding accuracy on the two sets of data. The results show that the system that is suggested outperforms current ones in terms of emotion detection accuracy. Furthermore, the results obtained on the one dataset with feature images are encouraging, demonstrating the necessary expertise for practical applications. The work's drawback is that precision still needs to be improved. The shortcoming of the work is that accuracy still needs to be improved.

Shahzad et al. [33] applied CNN for emotion classification. On both the AlexNet and VGG-16 architectures, simulation results show that support vector machine (SVM) and ensemble classifiers surpass the SoftMax classifier. The study reveals the feasibility of leveraging the positive characteristics of CNNs and other machine learning (ML) methods to improve performance in emotion recognition applications. The authors present an approach for analyzing different techniques to increase image classification performance by isolating learned features from CNNs that have already been trained and employing a range of classifiers. The drawback of the work is that its accuracy is low. Wang et al. [34] developed a system for multifaceted emotion detection through facial expressions as a small-scale learning difficulty and provided self-cure relation networks (SCRNet). To address the label noise issue, a prototype model was kept in auxiliary storage and was used to solve the challenge of noise in labels throughout the meta-training stage. Simulation on real and synthetic noise datasets indicates the technique's practicality. The downside of the work is that the proposed system's accuracy is low.

Zhu et al. [35] investigated the association between face recognition of emotions and behavioral factors. On this premise, a facial emotion detection model is created by expanding the layers of the convolutional neural network (CNN) and merging CNN with various neural networks for facial emotion detection. Following the preprocessing of images of faces and the tuning of important variables, the technique's effectiveness is evaluated. Image preparation and parameter adjustment enhance this method's recognition accuracy, and there is no ill-fitting. The work offers statistical references and guidance for studying the emotional traits of adolescents who engage in criminal activity. Farhoumandi et al. [36] proposed a new diagnostic technique that employs machine learning models built around face emotion detection test scores. In a study with a cross-sectional approach, fifty-five learners were chosen from a university based on the requirements for inclusion and exclusion as well as their Toronto Alexithymia Scale (TAS-20) ratings. To predict schizophrenia, two machine learning classifiers were developed employing K-fold cross-validation, and the model's effectiveness was measured using different metrics. After choosing features and maximizing efficiency, the models achieved a maximum accuracy of 81.8%. The findings indicated that ML models could accurately discriminate schizophrenia and identify which variables were most relevant for predicting schizophrenia. Results demonstrate that ML models are capable of accurately diagnosing schizophrenia. They may also serve as a medical tool to assist doctors in making diagnoses and timely identifying the illness. The performance of the system is relatively low. Table 1 summarises the related works, the contributions, and research gap that led to the present research.

Every study significantly outperformed earlier research on emotion recognition, but a straightforward strategy for identifying crucial facial regions for emotion detection is still lacking. This paper used CNN-10 with data augmentation and ViT models that concentrate on important face features to overcome the challenges inherent in the abovementioned works.

3. Methodology

3.1. Proposed Models. This section describes the high-level block diagram of the CNN-10, Vision Transformer (ViT), InceptionV3, and VGG19 models used for facial expression analysis in this paper. The diagram includes key components such as the data set, cross-validation, and data augmentation techniques used in this work. The following steps outline the processes involved in this work, as shown in Figure 1.

(1) Dataset: A carefully curated dataset of human facial expression images is collected. This dataset contains labeled images representing various facial expressions such as happiness, sadness, anger, and more. These images serve as the input data for training and evaluating the CNN models.

(2) Data Augmentation: To enhance the diversity and robustness of the dataset, data augmentation techniques are applied. These techniques involve applying transformations such as rotation, scaling, flipping, and cropping to the existing images. Data augmentation helps increase the variability of the training data and improves CNN's ability to generalize to unseen images.

(3) Cross-Validation: To assess the performance and generalization capabilities of the CNN models, cross-validation is employed. The dataset is divided into multiple folds, or subsets. The training and evaluation process is repeated several times, with each fold serving as the validation set once. This approach ensures a comprehensive evaluation of the CNN models' performance across different subsets of the data.

(4) CNN Architecture: The high-level block diagram incorporates four distinct CNN models: CNN-10, vision transformer (ViT), InceptionV3, and VGG19. These models are specifically chosen for their effectiveness in image analysis tasks, including facial expression recognition.

   (a) CNN-10: This is a CNN architecture with ten convolutional layers. It is designed to capture hierarchical features from input images, gradually extracting more abstract representations.

Table 1: Summary of all the related works with other algorithms and contributions.

Ref	Algorithms proposed	Contributions	Research gap
Fan and Tjahjadi [26]	Deep learning and handcrafted features	The proposed method achieved a 92.5% facial recognition rate	Relatively small dataset was used. Efficacy of the method was not compared with any well-known high-performing deep learning algorithm
Minaee et al. [27]	Deep emotion using convolutional network	The proposed method achieved a 99.0% facial recognition rate	Inability of the proposed method to adequately cope with the imbalanced nature of different emotion classes in the dataset used
Zhao et al. [28]	Deep learning	The proposed method achieved a 90.95% facial recognition rate	The proposed work cannot be classified directly, and it is necessary to increase the accuracy of the approach
Abdulrahman and Eleyan [29]	Support vector machines with principal component analysis (PCA) and local binary pattern (LBP) algorithms	The proposed method achieved 87% and 77% facial recognition rates	Data cannot capture the main feature on a very large scale
Rescigno et al. [30]	Transfer learning	Achieved RMSE of 0.09 and RMSE of 0.1 for valence and arousal, respectively	The pretrained net performed very poorly and so more difficult to be generalized
Jain et al. [31]	Extended deep neural networks	The proposed method achieved 95% facial recognition rate	Accuracy is pretty low and needs further improvement
Akhand et al. [32]	Transfer learning in the deep CNN	Attained accuracies of 96.51% and 99.52%, respectively, with DenseNet-161 on KDEF and JAFFE	Accuracy still needs to be improved
Shahzad et al. [33]	Deep convolutional neural networks (DCNNs)	Improved accuracies between 7% and 9%	Accuracy is low
Wang et al. [34]	Label noises self-cure relation networks	A prototype model was kept in auxiliary storage and used to solve the challenge of noise in labels throughout the meta-training stage	Low accuracy
Zhu et al. [35]	CNN and neural networks	Attained an average detection rate of 88.16%. Offers statistical references and guidance for studying the emotional traits of adolescents who engage in criminal activity	Performance is comparatively low
Farhoumandi et al. [36]	Machine learning	Maximum accuracy range of 81.8%	Moderately low accuracy

(b) Vision Transformer (VIT): ViT is a transformer-based architecture adapted for image analysis. It employs self-attention mechanisms to capture spatial relationships and global context within the image.
(c) InceptionV3: InceptionV3 is a deep CNN model known for its efficiency and accuracy. It consists of multiple parallel convolutional layers with different filter sizes, allowing for the extraction of both local and global image features.
(d) VGG19: VGG19 is a deep CNN model with 19 layers, known for its simplicity and strong performance. It utilizes multiple convolutional layers followed by fully connected layers, enabling effective feature extraction.

(5) Facial Expression Analysis: The trained CNN models (CNN-10, ViT, InceptionV3, and VGG19) are utilized to analyze facial expressions in the input images. These models have learned to classify facial expressions into different categories based on the features extracted from the images. The output of the CNN models provides the predicted facial expression for a given input image.

By adopting this high-level block diagram, incorporating CNN-10, ViT, InceptionV3, and VGG19 models, along with the dataset, cross-validation, and data augmentation techniques, researchers can develop a robust facial expression analysis system. This approach enhances the CNN models' ability to accurately classify and recognize facial expressions in human images, which is valuable for various applications such as emotion recognition, human-computer interaction, and psychological research.

3.2. Dataset. This work makes use of three datasets: CK+ [37], FER-2013 [38], and JAFFE [39] dataset. The CK+ dataset was gotten from the Kaggle website [37]. Figure 2 depicts the facial recognition of imbalanced dataset. The original dataset comprises 732 images with unequal instances of happiness (207), disgust (177), anger (135), sadness (84), fear (75) and contempt (54) as shown in Figure 3.

3.3. Data Augmentation. A total of 60,000 75×75 augmented images were created by carefully cropping each image in the collection into numerous 75×75 pieces. Each of the images contained multiple images of various facial expressions [40, 41]. The training and validation set consists of 8,000 each of anger, contempt, disgust, fear, happiness, and sadness, as shown in Figure 4. Also, the test set consists of 2,000 each of anger, contempt, disgust, fear, happiness, and sadness, as depicted in Figure 5. The dataset was randomly divided into training data (64%), validation data (20%), and test (16) for each facial recognition, as illustrated in Figures 4 and 5. In the meantime, we made sure that the images in the two datasets did not overlap. We supplemented the training, validation, and test data by setting the rotation range as 15, the width shift range as 0.2, the height shift range as 0.2, the shear range as 0.2, and the zoom range as 0.2. This technique enables us to increase the training, validation, and test size of the image.

3.4. Cross-Validation. Hold-out cross-validation, also known as simple cross-validation, was applied in this work. Hold-out cross-validation is a technique used to assess the performance of a deep learning model. It involves splitting the available dataset into two disjoint subsets: a training set and a validation/test set. Hold-out cross-validation is a straightforward and commonly used method for evaluating deep learning models. It provides an estimate of the model's performance on unseen data by simulating the real-world scenario of training on one set of data and evaluating on another. In this study, the dataset was split into a training, validation, and test set using a 64 : 20 : 16 split, respectively. The random_state parameter ensures the reproducibility of the splits, as shown in Figure 6.

3.5. Deep Learning Models

3.5.1. Convolution Neural Network (CNN-10). The CNN-10 architecture, which was used in this paper to classify facial expressions, consists of a few building blocks, including two convolution layers, leaky-ReLU, batch normalization, max pooling, two drop-out, flatten, and two dense layers, as shown in Figure 7. As previously said, the purpose of the first CNN is to extract the most important characteristics and the various spatial scale representations from the input of facial expression. A filtering step plus the use of an activation function make up a convolution layer. The convolutional filter size is set to 32 and the kernel size is set to 33, and the number of filters used to define this layer corresponds to the depth of the output feature map that the convolutional layer produces. The weights for each pixel are represented by the filter values. Backpropagation, a typical procedure in the training of feed-forward neural networks, is used to randomly initialize and change these filter weights. The leaky ReLU layer, which is present in the third layer and is set to 0.5, allows us to deal with gradient mortality and eliminate gradient problems that may easily arise during the backpropagation process while conducting neural activation between layers. The fourth layer in this study is batch normalization.

Before transferring the output of the third layer onto the input of the next fifth layer, batch normalization helps to normalize the output. The fifth layer is the most well-known type of pooling operation, called max pooling. Maximum pooling is set to a pooling size of 2 by 2. To avoid overfitting, the dropout layers were used in the sixth and ninth layers. The flattening layer, which is the seventh layer, allows us to convert the multidimensional input to one dimension. Dense or fully connected layers are the eighth and tenth layers. In the ultimate fully connected layer, the number of output nodes typically corresponds to the number of groups, which is set at 6.

3.5.2. Vision Transformer (ViT). Unlike CNN-based models for image classification tasks, ViT design [42], which is fully based on transformer architecture, exhibited outstanding

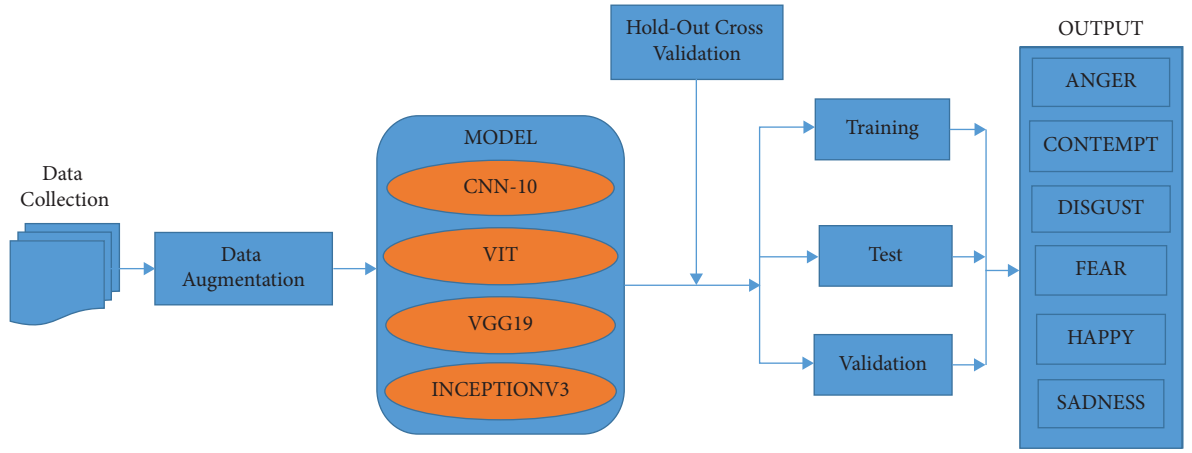

Figure 1: Block diagram of facial emotion recognition.

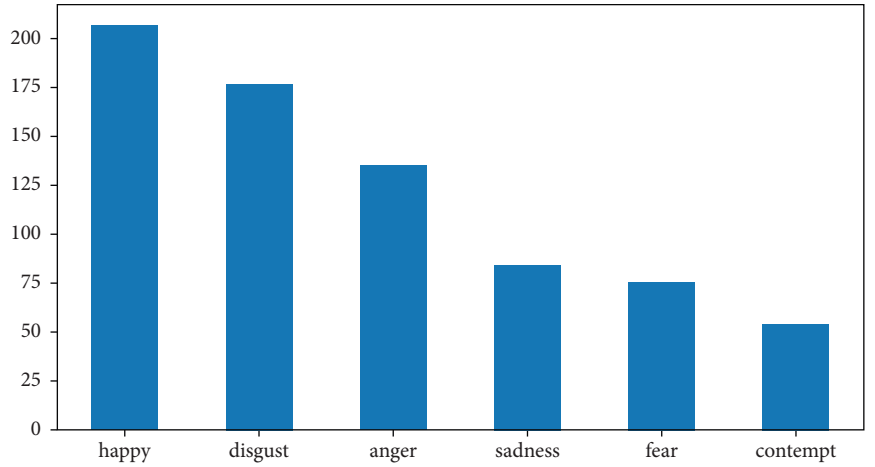

Figure 2: Facial recognition of an imbalanced dataset.

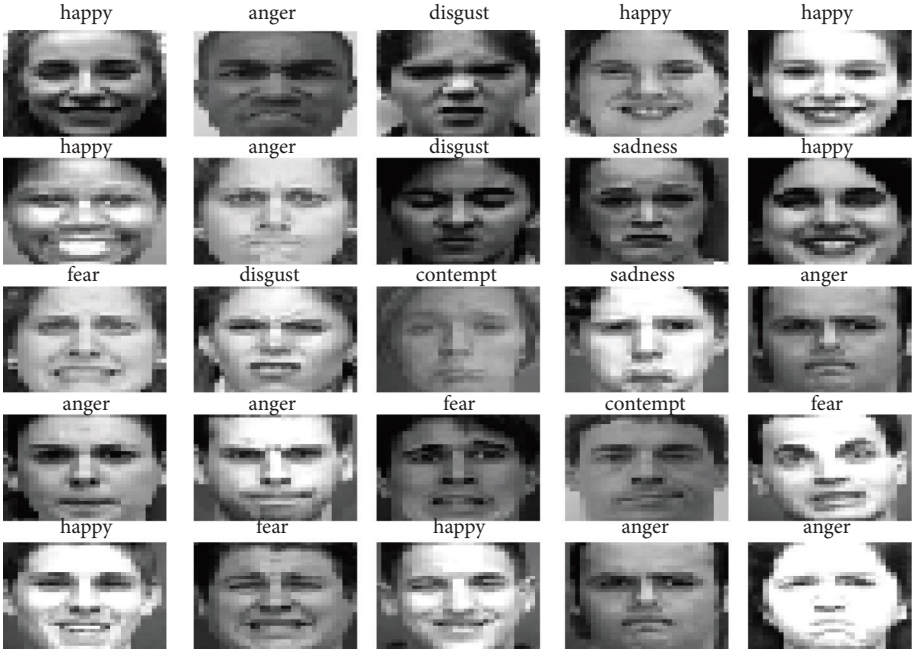

Figure 3: Sample of facial expression of a human being.

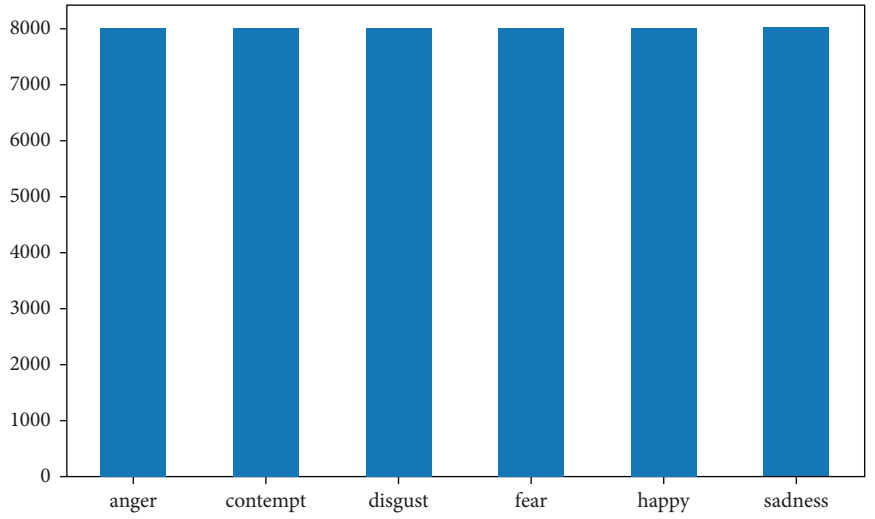

Figure 4: Training and validation set of facial recognition after data augmentation.

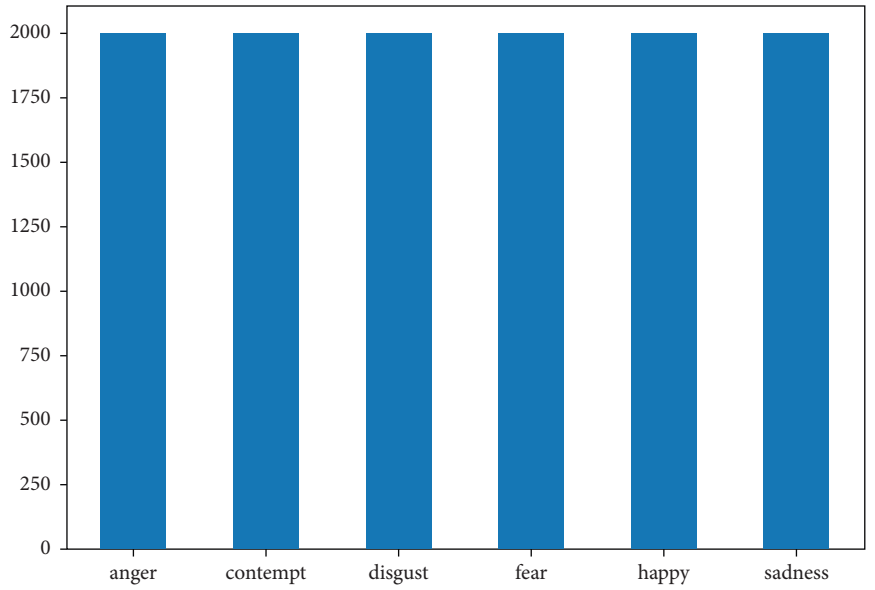

Figure 5: Test set of facial recognition after data augmentation.

accuracy. ViT captures long-range relationships between input sequences through a self-attention mechanism. ViT helps to classify images using the transformer model. It separates the input picture into a few linearly projected patches, uses learnable positional embedding to determine the order of the patches, and then uses a transformer encoder and multilayer perceptron to get the final classification. The input image is separated into nonoverlapping patches in the first section since a conventional transformer requires a 1D token sequence as input. Typically, images are in 2D format; therefore, in this research, images with dimensions of 75 in height, 75 in weight, and 3 in channels are taken into consideration, with a 14 by 14 image patch size. The elements per patch are set at 588, and there are 25 patches per image. The mathematical representations of the multiheaded self-attention (MSA), multilayer perceptron (MLP), and layer norm (LN) are applied before every block, respectively. These are represented in equations (1)–(4):

$$z_o = \left[x_{\text{class}}; x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + E_{\text{pos}}, \tag{1}$$

$$z_l' = \text{MSA}\left(\text{LN}\left(z_{l-1}\right) + z_{l-1}, \right. \tag{2}$$

$$z_l = \text{MLP}\left(\text{LN}\left(z_l'\right)\right) + z_l', \tag{3}$$

$$y = \text{LN}\left(z_L^o\right). \tag{4}$$

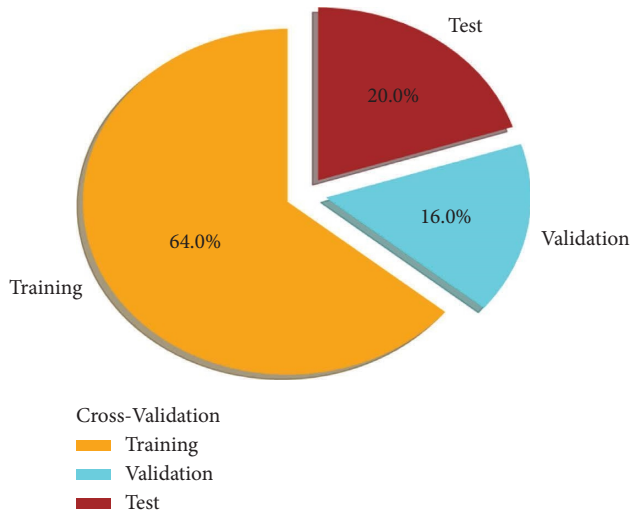

Figure 6: Cross-validation of facial expression.

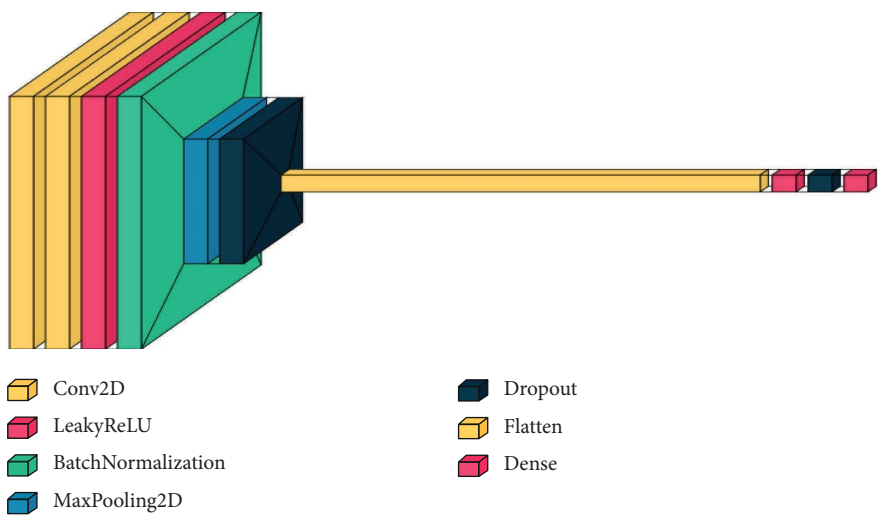

Figure 7: Proposed model architecture.

3.5.3. InceptionV3. The Inception-v3 model is pretrained for our investigation. Inception-v3 [43] has three different types of Inception modules: Inception A, Inception B, and Inception C. The well-designed convolution modules used in Inception may both produce distinguishing characteristics and minimize the number of parameters. Each Inception module is made up of parallel convolutional and pooling layers. The Inception modules employ small convolutional layers like 3 by 3, 1 by 3, 3 by 1, and 1 by 1 layers to minimize the number of parameters. There are 1,000 classes in the output of the original Inception-v3 network, but we only had six, namely, contempt, disgust, fear, happiness, and sadness. As a result, we adjusted the last layer output channel count from 1,000 to 6.

3.5.4. VGG19. VGG19 is a convolutional neural network introduced by [44] that comprises of 19 layers, 16 convolution layers, and 3 fully connected layers to categorize images into 1000 item categories. During training, validating, and testing VGG19, we resized the images to 75 by 75, but we did not modify the number of channels, only the size of the feature maps created during the operation. The last layer of VGG19 was omitted because it was only used for imageNet completion. Then, towards the conclusion of the VGG19 modules, we added global average pooling and fully connected layers to allow us to use the pretrained model and fine-tune the parameters for our particular purpose. The last step involved adding a softmax layer as a classifier that outputs probabilities for each class. The class with the maximum likelihood was selected as the projected class. A total of 1,000 classes were produced by the original VGG19 network, but we only had six, including disgust, fear, happiness, and sadness. As a result, we adjusted the last layer's output channel count from 1,000 to 6.

3.6. Evaluation Metrics. Different assessment measures including precision, recall, and $F1$-score are used to assess the performance of system models.

3.6.1. Precision. It counts the number of times a model predicts positively accurately out of all positively predicted outcomes. This is depicted in the following equation:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{5}$$

3.6.2. Recall/Sensitivity. It counts the number of times a model predicts a label positive from an overall positive class accurately. This is represented in the following equation:

$$\frac{\text{Recall}}{\text{Sensitivity}} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{6}$$

3.6.3. F1 Score. It combines recollection, balance, and both precision and recall. The weakest model has an $F1$ score of 0, while the best gets 1. A higher $F1$ score indicates that the model has fewer false positives and negatives. This is shown in the following equation:

$$F1 - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{7}$$

4. Results and Discussion

This section presents the results of the experiments conducted using selected deep learning algorithms on three different datasets. Figure 2 shows the graphic representations of the facial recognition datasets used to test the performance of the suggested framework.

4.1. Experimental Setting. Python (version 3.9) is used to implement the suggested approach in a Jupyter notebook environment. During experimentation, the well-known deep learning frameworks, namely, TensorFlow (2.10.0 version), Numpy (1.23.1), Pandas (1.5.0), seaborn (0.12.0), and scipy (1.9.1) are utilized. Table 2 gives the details of the parameters.

The standard evaluation measures mentioned above were used to assess the efficacy of the approach proposed in this work. The performance evaluation of the training, validation, and test sets is presented in Table 3.

The proposed model (CNN-10) has the highest training, validation, and test accuracy with 99.95%, 99.91%, and 99.80%, respectively. Table 4 is the loss evaluation of the training, validation, and test scores with 0.0016, 0.0028, and 0.0267 which is relatively low compared to other algorithms considered in this study.

Table 5 depicts the precision, recall, and $F1$-Score of VGG19 pretrained model. The results demonstrate that disgust, fear, happiness, and sadness classes have the lowest measurements among the classification labels. Anger has a precision of 0.22, recall of 1.00, and $F1$-score of 0.36; class contempt has a precision of 0.67, recall of 1.00, and $F1$-score of 0.80.

Table 2: Parameter settings.

Parameter	Value
Filter	32, 16
Kernel_size	3
Padding	Same
Activation	ReLU, SoftMax
LeakyReLU	0.5
Pool_size	2
Dropout	0.25, 0.2
Dense	100
Loss	Categorical_crossentropy
Optimizer	Adam

Table 3: Accuracy score of cross-validation of the models for CK+.

Models	Training (%)	Validation (%)	Test (%)
VGG19	15.47	15.52	0.00
INCEPTIONV3	21.61	19.53	1.25
ViT	16.51	15.21	0.00
CNN-10	99.95	99.91	99.80

The precision, recall, and $F1$-score of the class using INCEPTIONV3 values are depicted in Table 6. The outputs prove that contempt and happiness have the least values among the classification labels. Anger has a precision of 0.64, recall of 0.02, and $F1$-score of 0.04; disgust has a precision of 0.78, recall of 0.04, and $F1$-score of 0.08 while sadness has a precision of 0.66, recall of 0.09, and $F1$-score of 0.16.

From Table 6, INCEPTIONV3 has a difficulty recognizing contempt and happiness emotions for the CK+ dataset. The model has a relatively average performance in recognizing disgust, sadness, and anger emotions from the CK+ dataset. Based on this average performance, it cannot be recommended that the model be adopted for facial emotion applications as it struggles to recognize some emotions.

In Table 7, disgust is the only class having the highest measurement with precision of 0.17, recall of 1.00, and $F1$-Score of 0.29.

Results in Table 7 indicated that ViT has a very poor performance as it failed to recognize all the motions represented in the CK+ dataset except disgust. All the other emotions, such as anger, contempt, fear, happiness, and sadness, in the CK+ dataset was not detected by the ViT model. Based on this abysmal performance, it can be suggested that the model cannot be used for real-world facial emotion applications as it finds it difficult to recognize many of the emotions in the CK+ dataset.

CNN performs excellently well in Table 8, and all the classes have the highest value of 1.00 in both precision, recall, and $F1$-Score.

The statistical results depicted in Table 8 show that CNN-10 performs excellently as it successfully recognises all the motions represented in the CK+ dataset without any exception. All the emotions, such as anger, contempt, disgust, fear, happiness, and sadness, in the CK+ dataset were

Table 4: Loss score of cross-validation of the model for CK+.

Models	Training	Validation	Test
VGG19	1.7940	1.7938	2.1231
INCEPTIONV3	302.0092	323.2202	410.5343
ViT	3.7419	3.5340	1.2215
CNN-10	0.0016	0.0028	0.0267

Table 5: Precision, recall, and *F*1-score of VGG19 for CK+.

Models	Facial-recognition	Precision	Recall	*F*1-score
VGG19	Anger	0.22	1.00	0.36
	Contempt	0.67	1.00	0.80
	Disgust	0.00	0.00	0.00
	Fear	0.00	0.00	0.00
	Happiness	0.00	0.00	0.00
	Sadness	0.00	0.00	0.00

Table 6: Precision, recall, and *F*1-score of INCEPTIONV3 for CK+.

Models	Facial-recognition	Precision	Recall	*F*1-score
INCEPTIONV3	Anger	0.64	0.02	0.04
	Contempt	0.00	0.00	0.00
	Disgust	0.78	0.04	0.08
	Fear	0.57	0.11	0.18
	Happiness	0.18	1.00	0.30
	Sadness	0.66	0.09	0.16

Table 7: Precision, recall, and *F*1-score of ViT for CK+.

Models	Facial-recognition	Precision	Recall	*F*1-score
ViT	Anger	0.00	0.00	0.00
	Contempt	0.00	0.00	0.00
	Disgust	0.17	1.00	0.29
	Fear	0.00	0.00	0.00
	Happiness	0.00	0.00	0.00
	Sadness	0.00	0.00	0.00

Table 8: Precision, recall, and *F*1-score of CNN-10 for CK+.

Models	Facial-recognition	Precision	Recall	*F*1-score
CNN-10	Anger	1.00	1.00	1.00
	Contempt	1.00	1.00	1.00
	Disgust	1.00	1.00	1.00
	Fear	1.00	1.00	1.00
	Happiness	1.00	1.00	1.00
	Sadness	1.00	1.00	1.00

perfectly detected by the CNN-10 model. Based on this outstanding performance, it can be concluded that the CNN-10 model will be a promising technique for detecting facial emotions in real-world applications, as it perfectly recognises all the emotions in the CK+ dataset.

Depicted in Figure 8 are patches of facial emotional expression. It shows the patches obtained from ViT, which shows that it can detect eyebrows, eyes, nose, mouth, nostrils, and the directions in which a human is facing.

The results of accuracy, macroaverage, and weighted average for each deep learning algorithm are represented in Figure 9.

The results show that the proposed CNN-10 model outperforms other models with a score of 99.99% for accuracy, macroaverage, and weighted-average. The VGG19 model shows its efficacy with an accuracy score of 33%, a macroaverage of 19%, and a weighted average of 19%. For INCEPTIONV3 models, the accuracy is 21%, the macroaverage is 13%, and the weighted average is 13%. The ViT models show the lowest accuracy, with an accuracy of 17%, a macroaverage of 5%, and a weighted average of 5%. As shown in the results, the proposed models perform better than other models compared to this work.

Figures 10–13 are the confusion matrix of VGG19, INCEPTIONV3, ViT, and CNN-10.

Figure 10 shows the confusion matrix of VGG19, which provides valuable insights into the performance of a classification model for different classes. The confusion matrix shows the predicted and actual class labels for a multiclass classification problem involving six emotions: anger, contempt, disgust, fear, happiness, and sadness. For anger, out of 1999 instances of the anger class, the model correctly predicted all of them as anger (true positives). There were no instances misclassified as any other class (false negatives). For contempt, similarly, all 2000 instances of the contempt class were correctly classified as contempt (true positives), with no misclassifications. For disgust, out of 1995 instances of the disgust class, the model correctly classified 995 as disgust (true positives). However, 1005 instances were misclassified as contempt (false negatives). For fear, all 2000 instances of the fear class were misclassified as anger (false negatives), indicating that the model failed to recognize any instance of fear. For happiness, all 2000 instances of the Happiness class were misclassified as anger (false negatives), suggesting that the model did not correctly identify any instance of happiness. For sadness, all 2000 instances of the class were misclassified as anger (false negatives), indicating that the model did not recognize any instances of sadness.

Figure 11 shows the confusion matrix, which represents the predicted and actual class labels for a multi-class classification problem using the InceptionV3 model. The figure consists of six emotions: anger, contempt, disgust, fear, happiness, and sadness. For anger, out of 2000 instances of the anger class, the model correctly classified 36 as anger (true positives). However, it misclassified 1949 instances as happiness and 13 instances as sadness (false negatives). For contempt, the model did not predict any instances of contempt for the given class. It misclassified all 2000 instances as happiness (false negatives). For disgust, out of 2000 instances of the disgust class, the model correctly classified 82 as disgust and 1766 as happiness (true positives). However, it misclassified 20 instances as anger, 113 instances as fear, and 19 instances as sadness (false negatives). For fear, the model correctly classified 216 instances of fear out of 2000 (true positives). However, it misclassified 1748 instances as happiness and 36 instances as sadness (false negatives). For happiness, the model correctly classified all 2000 instances as happiness (true positives) with no

Figure 8: Patches of facial emotional expression.

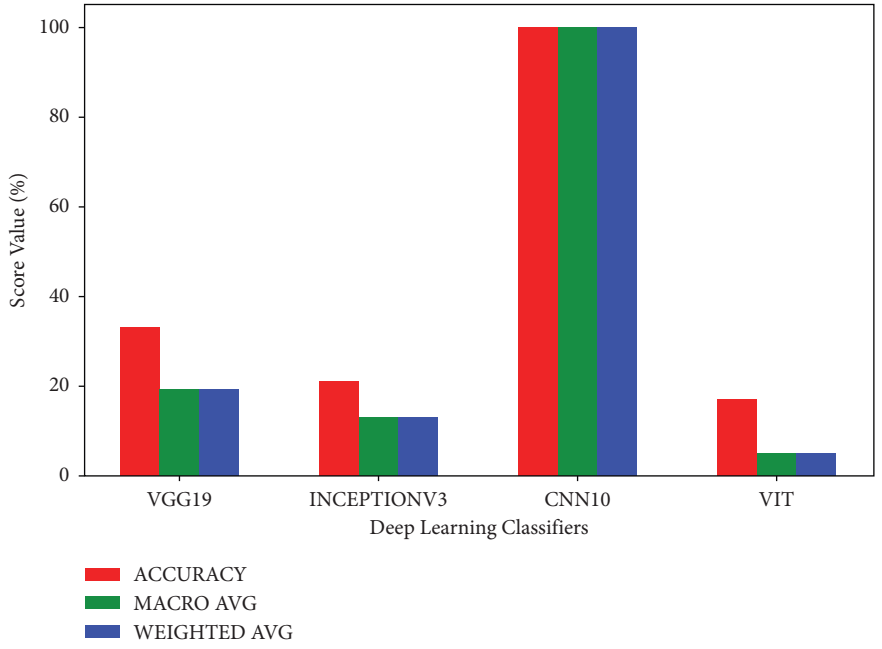

Figure 9: Overall accuracy, macroaverage, and weighted-average of the models.

misclassifications. Sadness: out of 2000 instances of the sadness class, the model correctly classified 23 as sadness and 180 as happiness (true positives). However, it misclassified 45 instances as disgust (false negatives).

Figure 12 shows the confusion matrix, which represents the predicted and actual class labels for a multiclass classification problem using the vision transformer (ViT) model. The figure consists of six emotions: anger, contempt, disgust, fear, happiness, and sadness. Since the entire column contains zeros except for the diagonal elements, this indicates that the ViT model has classified all instances as belonging to the disgust class. For anger, the model classified all instances (2000) as disgust, resulting in zero instances correctly classified as anger. Contempt: like anger, the model classified all instances (2000) as disgust, resulting in zero instances correctly classified as contempt. Disgust: the model classified all instances (2000) as disgust, correctly classifying them as the true positive. Fear: once again, the model classified all instances (2000) as disgust, resulting in zero instances correctly classified as fear. Happiness: the model classified all instances (2000) as disgust, resulting in zero instances correctly classified as happiness. Sadness: like the other classes, the model classified all instances (2000) as disgust, resulting in zero instances correctly classified as sadness.

Figure 13 shows the confusion matrix, which represents the predicted and actual class labels for a multiclass classification problem using the CNN-10 model. The figure consists of six emotions: anger, contempt, disgust, fear, happiness, and sadness. Anger: the model correctly classified all instances (2000) as anger, resulting in a true positive. Contempt: the model classified 1994 instances correctly as contempt, with 3 instances incorrectly classified as anger, 2 instances incorrectly classified as fear, and 1 instance incorrectly classified as happiness. Disgust: the model correctly

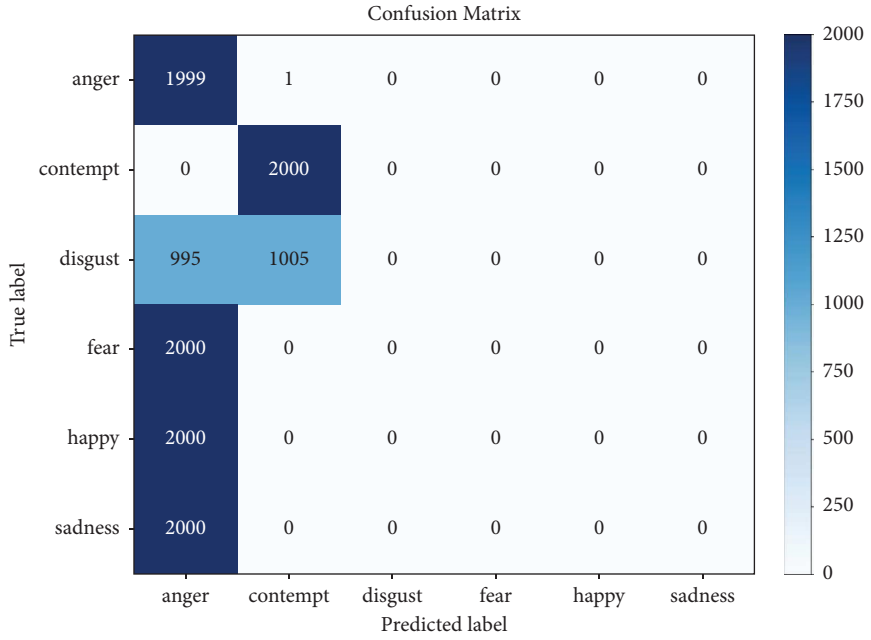

Figure 10: Confusion matrix of VGG19.

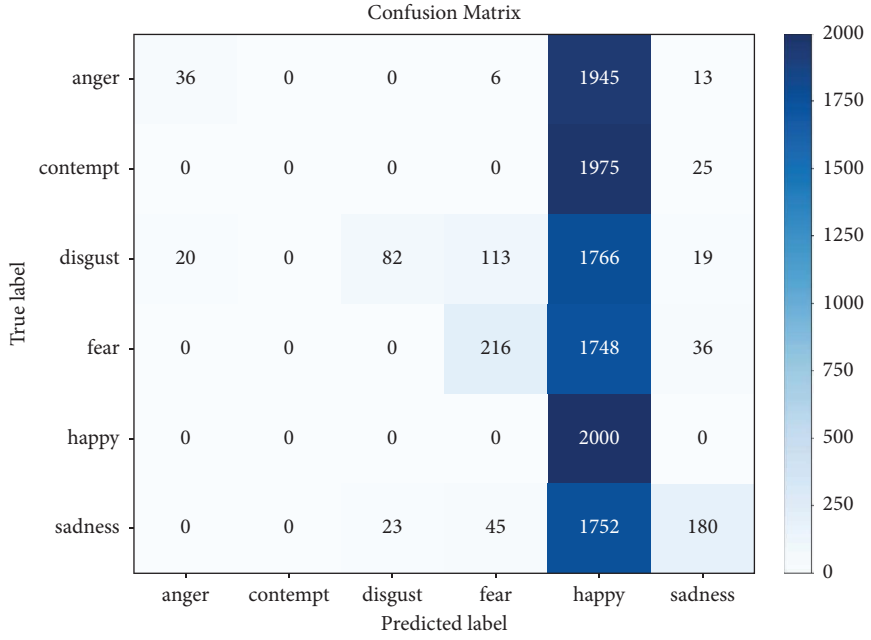

Figure 11: Confusion matrix of InceptionV3.

classified 1995 instances as disgust, with 2 instances incorrectly classified as anger and 3 instances incorrectly classified as fear. Fear: the model correctly classified 1995 instances as fear, with 3 instances incorrectly classified as anger and 1 instance incorrectly classified as sadness. Happiness: the model correctly classified 1997 instances as happiness, with 1 instance incorrectly classified as contempt, 1 instance incorrectly classified as disgust, and 1 instance incorrectly classified as sadness. Sadness: The model correctly classified 1997 instances as sadness, with 2 instances incorrectly classified as disgust and 1 instance incorrectly classified as happiness.

Figure 14 illustrates the FER-2013 facial recognition dataset, encompassing a diverse array of emotional expressions. This dataset is categorized into seven distinct emotion classes, each characterized by a varying number of samples: “Surprise” with 3,171 samples, “Sadness” featuring 4,830 samples, “Neutral” comprising 4,965 samples,

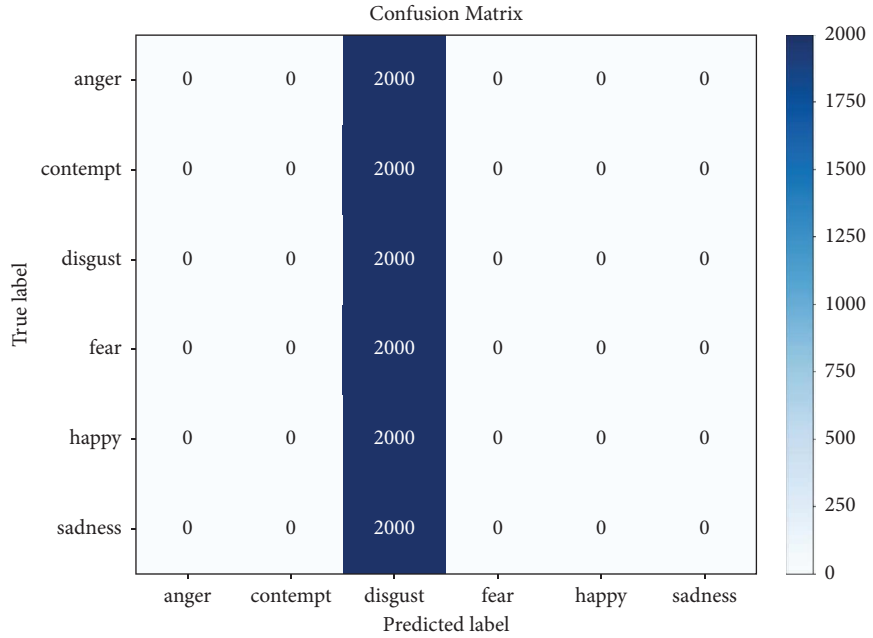

Figure 12: Confusion matrix of ViT.

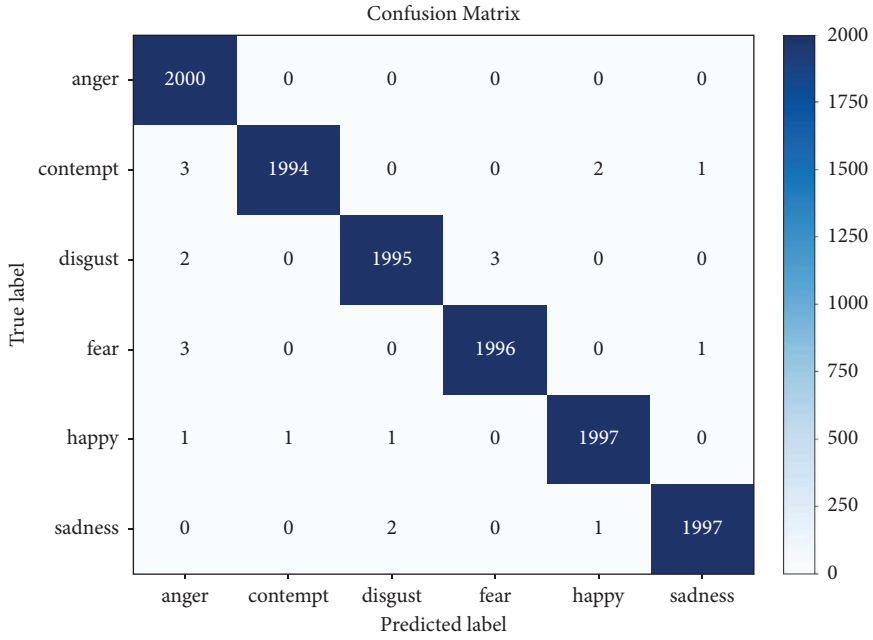

Figure 13: Confusion matrix of CNN-10.

"Happiness" containing 7,215 samples, "Fear" represented by 4,097 samples, "Disgust" observed in 436 samples, and "Anger" conveyed through 3,995 samples. In Figure 15, we present the dataset sourced from the Japanese Female Facial Expression Database (JAFFED) for facial recognition. This dataset encompasses a diverse set of emotional categories, each represented by a specific number of samples. Specifically, the dataset comprises 30 samples denoting expressions of "Surprise," 31 samples capturing "Sadness," 30 samples conveying "Neutral" emotional states, 31 samples exemplifying "Happiness," 32 samples reflecting "Fear," 29 samples depicting "Disgust," and 30 samples representing "Anger".

4.2. Critical Analysis and Discussion. This section provides a brief critical analysis and discussion of the results obtained from the experiments conducted in this work. For the

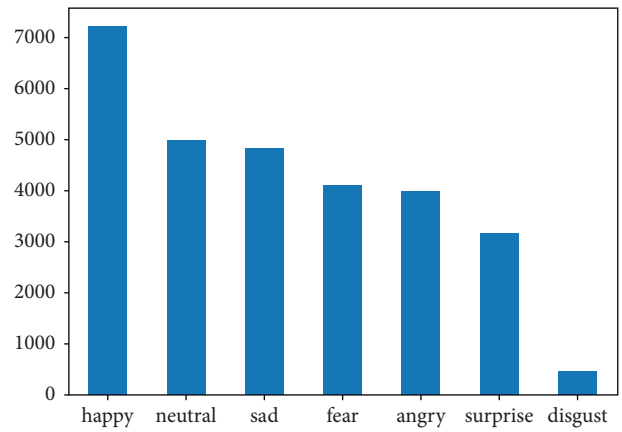

Figure 14: Dataset of FER-2013 facial recognition.

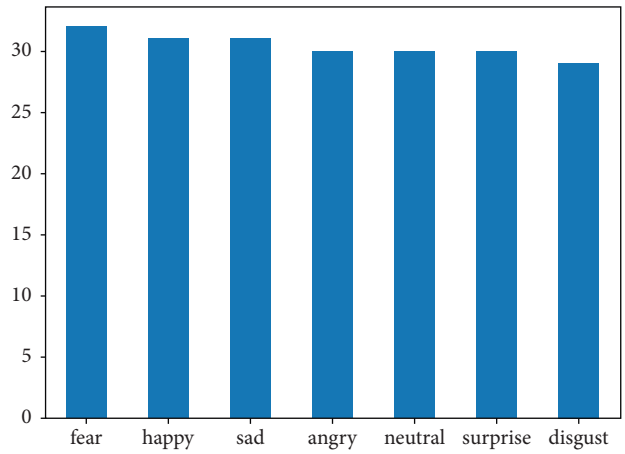

Figure 15: Dataset of JAFFE of facial recognition.

VGG19 model, one can observe that the model performs well for the anger and contempt classes, correctly predicting all instances. However, it struggles with the disgust, fear, happiness, and sadness classes, misclassifying them as anger. This suggests that the model may have difficulty distinguishing between these emotions or may not have learned their distinctive features effectively. For the InceptionV3 model, it can be observed that the InceptionV3 model performs well for the happiness class, correctly predicting all instances. However, it struggles with other emotions such as anger, contempt, disgust, fear, and sadness, misclassifying them as happiness. This suggests that the model may have difficulty distinguishing between these emotions or may not have learned their distinctive features effectively. For ViT model based on the results, it appears that the ViT model is incorrectly classifying all instances as disgust, regardless of the true emotion. This suggests that the model may not have learned the distinguishing features of the different emotions and is not able to generalize well enough to classify them correctly. Taking a critical look at the proposed CNN-10 model, from the results, it is evident that the CNN-10 model has performed well overall, correctly classifying most instances into their respective classes. However, there are some instances where misclassifications occurred, particularly between similar emotions like anger and contempt and disgust and fear.

Tables 9–11 present a performance comparison of all the models on the CK+, FER-2013, and JAFFE datasets, respectively.

Tables 9–11 present comparisons between the outputs of several approaches on the CK+, FER-2013, and JAFFE datasets. The CNN-10 model, as described in this study, had greater performance when compared to other modern methodologies that were examined. The CNN-10 model demonstrates a notable level of accuracy, indicating its proficiency in reliably identifying and categorising facial expressions across all datasets included in this research. The following tables depict the accuracy scores attained by different models when applied to the CK+, FER-2013, and JAFFE datasets. The results from the experiment demonstrate that CNN-10 exhibits superior performance compared to the other models employed in this study, as evidenced by measures including accuracy, recall, precision, and F1 score. The rationale behind utilising the CNN model lies in its significant computational capabilities, which were effectively used during the picture processing phase. Therefore, it can be concluded that the CNN model being offered has exceptional achievement compared to the other models that are currently being considered.

Table 12 presents an evaluation analysis comparing the proposed approach with several previous studies conducted in the domain of face emotion recognition. The table presents a comparison between the proposed strategy and the current contemporary approach in terms of accuracy. The present study compares the experimental results produced in this research with those acquired from several other recent techniques [27, 45–49, 51, 52]. In the scientific literature, scholars employ a diverse range of methods for the purpose of face emotion detection, many of which are rooted in the foundation of deep neural networks. Our model demonstrates a higher level of accuracy compared to the findings reported in other studies.

The suggested configuration of convolutional neural networks (CNNs) is considered a significant aspect that contributes to the enhanced results. The primary objective of deep neural networks is to identify the most effective hyperparameters that can expedite the learning process and minimize the loss when the model reaches convergence. Certain crucial hyperparameters, among them the inclusion of a hidden layer, the choice of activation function, the learning rate, and the number of iterations, have the potential to enhance the effectiveness of the model. Impulse is an additional crucial factor that plays a significant role in improving the accuracy of the model. Additional parameters, such as mini-batch sizes and regularisation, serve a vital role in the functioning of neural networks with deep layers.

4.3. Strength, Limitation, and Significance of the Work. One of the proposed CNN-10 model's strengths is excellent feature extraction. CNN-10 is ideally suited to capture spatial features in images, which makes it helpful for obtaining selective facial aspects associated with emotions.

Table 9: Performance comparison for CK+.

Models	Accuracy	Precision	Recall	*F*1-score
VGG19	0.7751	0.76	0.80	0.80
INCEPTIONV3	0.7635	0.77	0.56	0.58
ViT	0.5211	0.54	0.57	0.59
CNN-10	0.9995	1.00	1.00	1.00

Table 10: Performance comparison for FER-2013.

Models	Accuracy	Precision	Recall	*F*1-score
VGG19	0.6641	0.67	0.66	0.63
INCEPTIONV3	0.6321	0.62	0.62	0.63
ViT	0.4742	0.49	0.49	0.49
CNN-10	0.8430	0.83	0.83	0.83

Table 11: Performance comparison for JAFFE.

Models	Accuracy	Precision	Recall	*F*1-score
VGG19	0.9221	0.93	0.93	0.95
INCEPTIONV3	0.9245	0.93	0.93	0.94
ViT	0.6543	0.63	0.63	0.64
CNN-10	0.9541	0.96	0.96	0.96

Table 12: Performances of selected models on CK+, FER-2013, and JAFFE datasets.

Datasets	Models	Accuracy (%)
CK+	Attentional CNN [27]	98.00
	IB-CNN [45]	95.10
	IACNN [46]	95.37
	DTAGN [47]	97.20
	VGG-19 [48]	99.47
	Proposed model	99.95
FER-2013	Attentional CNN [27]	70.02
	VGG + SVM [49]	66.31
	GoogleNet [50]	65.20
	VGG backbone [51]	75.00
	VGG-19 [48]	65.41
	Proposed model	84.30
JAFFE	Attentional CNN [27]	92.80
	LBP + ORG features [49]	88.50
	Deep features + HOG [50]	90.58
	CNN + SVM [51]	95.31
	VGG-19 [48]	99.47
	Proposed model	95.41

CNN-10's multilayer architecture enables it to dynamically learn and represent complicated patterns and features, which is useful for facial expression identification. A further benefit is that it is robust against changes in facial appearance. CNN-10 can manage differences in facial appearance, such as varying head orientations, dimensions, and lighting settings. CNN-10's convolutional layers are meant to detect local patterns that are resistant to movement, rotation, and scaling, which aids in consistently capturing facial expressions under varying settings. In addition, CNN-10 has a strong ability to learn hierarchical structures. CNN-10 can learn hierarchical representations by gradually blending fundamental features (for example, edges and corners) with high-level conceptual interpretations. Because of the hierarchical learning, the network can capture localized facial features and their spatial links, which can be useful for emotion recognition. CNN-10 also has the advantage of being able to manage massive data sets with a huge number of samples, leading to more extensive training and greater adaptation. Because big facial expression datasets, such as CK+, are available, CNN-10 has learned rich representations and enhanced its efficacy in the detection of facial emotion problems. Finally, the proposed method has the advantages of simultaneous processing and effectiveness. The CNN-10 architecture enables effective parallelism, allowing it to take advantage of GPU processing power and expedite training and inference. CNN-10 is thus well-suited for real-time or near-real-time applications requiring speedy and accurate emotion recognition.

Some of the limitations of this work include the limited spatial and temporal gathering of data. CNN-10 is designed for gathering spatial features and might find it difficult to adequately represent temporal changes in facial emotions. Emotions are frequently communicated through slight variations over time, and CNN-10 may not catch these temporal fluctuations efficiently, thus resulting in less accurate emotion recognition. Furthermore, the proposed approach places a restricted emphasis on the global perspective. CNN-10 receptive fields are mostly concerned with local spatial features. They might not accurately record the global context and interactions between various facial areas, which are critical for accurately interpreting facial expressions.

An additional problem is the shortage of data for some expressions. The sample size and broad range of facial expression data sets tend to be limited. Certain emotions, such as disgust or contempt, may be scarce in everyday situations, leading to a lack of data for building appropriate models for these emotions. This may influence CNN-10's ability to recognize less common or more subtle emotions. In addition, CNN-10 is difficult to interpret and explain. Because CNNs are frequently regarded as black-box models, it is difficult to comprehend and clarify how they make decisions. Discussions or understandings of which facial features or areas correspond to various emotion predictions may be required by facial emotion detection systems. CNN-10 does not have inbuilt comprehensibility, which limits its usefulness in some situations, like systems that require openness or accountability.

This work has several significant advantages to the society. Facial emotion recognition has the potential to be useful in mental health applications, assisting in the examination, surveillance, and therapeutic management of problems related to mental health. CNN-10 from the results of our experiments can identify indicators of sadness, anxiety, or other emotional states by studying facial expressions, allowing for early identification and management. The proposed work can be applied to comprehending human behavior in a variety of scenarios, including evaluating client satisfaction research, examining emotional responses

in social encounters, and examining facial expressions in forensic investigations. CNN-10 if properly deployed can supplement conventional techniques of behavior analysis by providing impartial and automated evaluations. In educational environments, facial expression detection can be useful for monitoring student participation, attentiveness, and emotional reactions during school sessions. The CNN-10 system can give educators instantaneous feedback, allowing them to modify instructional tactics to maximize the learning experience for students. CNN-10-based facial expression recognition can be incorporated into assistive technology to help handicapped people. It can, for example, provide emotion-based device control or improve interaction for those with speech problems by converting facial expressions into spoken or written output. Finally, in the discipline of robotics, face expression recognition can enhance interactions between humans and robots as well as robot behavioral intelligence. Robots can develop more spontaneous and interesting communications with humans by correctly detecting and reacting to the feelings that humans experience.

5. Conclusion

In this study, a CNN-10 approach for categorising facial emotional expression was presented, and the method was contrasted with others such as INCEPTIONV3, VGG19, and ViT. Particularly, the CNN-10 models categorization strategy is more accurate. In addition, INCEPTIONV3, VGG19, and VIT exhibit poor performance. CNN-10 can successfully increase the classification accuracy of facial emotion expressions, making it a reliable and effective computer-assisted diagnostic tool for identifying facial image data. Augmented images from the Kaggle dataset were utilized for classifier performance testing, validation, and training. The anger, contempt, disgust, fear, happiness, and sadness images that are present in the collection can all be easily recognized by CNN-10. The CK+ dataset having a 99.9% accuracy score, FER-2013 with a accuracy of 84.3%, and JAFFE with a accuracy of 95.4%. Our proposed technique, known as CNN-10 architecture, successfully recognises facial emotion expression. Therefore, future studies will concentrate on the selection of facial expression features via transfer learning. Other future research directions for this work will involve engaging in facial emotion recognition research towards depicting the mental health, health status, and internal wellbeing of individuals and how appropriate care and therapy could be facilitated towards improving the mental and overall health of individuals.

Data Availability

FER-2013 dataset for facial expression recognition is available at https://www.kaggle.com/datasets/msambare/fer2013 JAFFE: The Japanese female facial expression dataset is available at https://zenodo.org/record/3451524.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The APC was funded by the Virginia Tech University's research support to O.E. O.O.O received research funding from the Oppenheimer Memorial Trust (OMT) Foundation with grant number: OMT Ref. 21563/01; OOO was also supported with DAAD ClimapAfrica grant with grant number: ST32/91769426.

References

[1] S. Li and W. Deng, "Deep facial expression recognition: a survey," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2022.

[2] C. Han, L. Zhang, Y. Tang et al., "Understanding and improving channel attention for human activity recognition by temporal-aware and modality-aware embedding," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[3] H. Jung, S. Lee, S. Park et al., "Development of deep learning-based facial expression recognition system," in *Proceedings of the 2015 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)*, pp. 1–4, IEEE, Mokpo, South Korea, January 2015.

[4] O. Oluwagbemi and A. Jatto, "Implementation of a TCM-based computational health informatics diagnostic tool for Sub-Saharan African students," *Informatics in Medicine Unlocked*, vol. 14, pp. 43–58, 2019.

[5] O. Oluwagbemi, M. Keshinro, and C. Ayo, "Design and implementation of a secured census information management system," *Egyptian Computer Science Journal*, vol. 35, no. 1, pp. 1–11, 2011.

[6] O. Oluwagbemi, T. Ojutalayo, and N. Obinna, "Development of a secured information system to manage malaria related cases in southwestern region of Nigeria," *Egyptian Computer Science Journal*, vol. 34, no. 5, pp. 23–34, 2010.

[7] D. Cheng, L. Zhang, C. Bu, X. Wang, H. Wu, and A. Song, "ProtoHAR: prototype guided personalized federated learning for human activity recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 8, pp. 3900–3911, 2023.

[8] S. Xu, L. Zhang, Y. Tang, C. Han, H. Wu, and A. Song, "Channel attention for sensor-based activity recognition: embedding features into all frequencies in DCT domain," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–15, 2023.

[9] Y. Tang, L. Zhang, Q. Teng, F. Min, and A. Song, "Triple cross-domain attention on human activity recognition using wearable sensors," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 5, pp. 1167–1176, 2022.

[10] C. Darwin, *The Expression Of The Emotions In Man And Animals*, Murray, London, UK, 1948.

[11] V. A. Petrushin, "Emotion recognition in speech signal: experimental study, development, and application," in *Proceedings of the Sixth international conference on spoken language processing*, Beijing, China, October 2000.

[12] F. Jabr, *The Evolution of Emotion: Charles Darwin's Little-Known Psychology experiment*, Scientific American, New York, NY, USA, 2010.

[13] S. Dhall and P. Sethi, "Geometric and appearance feature analysis for facial expression recognition," *International Journal of Advances in Engineering & Technology*, vol. 5, no. 3, pp. 1–11, 2014.

[14] S. Mitsuyoshi and F. Ren, "Emotion recognition," *The journal of the Institute of Electrical Engineers of Japan*, vol. 125, no. 10, pp. 641–644, 2005.

[15] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, "Machine understanding of human behavior," *AI for Human Computing*, vol. 13, 2012.

[16] E. Hjelmas and B. Low, "Face detection: a survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.

[17] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[18] R. Jafri and H. R. Arabnia, "A survey of face recognition techniques," *Journal of Information Processing Systems*, vol. 5, no. 2, pp. 41–68, 2009.

[19] K. M. Malikovich, I. S. Z. Ugli, and D. L. O'ktamovna, "Problems in face recognition systems and their solving ways," in *Proceedings of the 2017 International Conference on Information Science and Communications Technologies (ICISCT)*, pp. 1–4, IEEE, Tashkent, Uzbekistan, November 2017.

[20] A. S. Dhavalikar and R. K. Kulkarni, "Face detection and facial expression recognition system," in *Proceedings of the 2014 International Conference on Electronics and Communication Systems (ICECS)*, pp. 1–7, IEEE, Coimbatore, India, February 2014.

[21] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *SN Applied Sciences*, vol. 2, no. 3, pp. 446–448, 2020.

[22] J. Brennan, *Facial Recognition: Defining Terms to Clarify Challenges*, vol. 13, Ada Lovelace Institute, London, UK, 2019.

[23] B. Peixoto, C. Michelassi, and A. Rocha, "Face liveness detection under bad illumination conditions," *Proceedings of the ICIP*, vol. 11, pp. 3557–3560, 2011.

[24] T. Kiran and T. Kushal, "Facial expression classification using support vector machine based on bidirectional local binary pattern histogram feature descriptor," in *Proceedings of the 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 115–120, IEEE, Shanghai, China, May 2016.

[25] J. H. Shah, Z. Chen, M. Sharif, M. Yasmin, and S. L. Fernandes, "A novel biomechanics-based approach for person re-identification by generating dense color sift salience features," *Journal of Mechanics in Medicine and Biology*, vol. 17, no. 07, Article ID 1740011, 2017.

[26] X. Fan and T. Tjahjadi, "Fusing dynamic deep learned features and handcrafted features for facial expression recognition," *Journal of Visual Communication and Image Representation*, vol. 65, Article ID 102659, 2019.

[27] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021.

[28] X. Zhao, X. Shi, and S. Zhang, "Facial expression recognition via deep learning," *IETE Technical Review*, vol. 32, no. 5, pp. 347–355, 2015.

[29] M. Abdulrahman and A. Eleyan, "Facial expression recognition using support vector machines," in *Proceedings of the 2015 23nd signal processing and communications applications conference (SIU)*, pp. 276–279, IEEE, Malatya, Turkey, May 2015.

[30] M. Rescigno, M. Spezialetti, and S. Rossi, "Personalized models for facial emotion recognition through transfer learning," *Multimedia Tools and Applications*, vol. 79, no. 47-48, pp. 35811–35828, 2020.

[31] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognition Letters*, vol. 120, pp. 69–74, 2019.

[32] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," *Electronics*, vol. 10, no. 9, p. 1036, 2021.

[33] H. M. Shahzad, S. M. Bhatti, A. Jaffar, S. Akram, M. Alhajlah, and A. Mahmood, "Hybrid facial emotion recognition using CNN-based features," *Applied Sciences*, vol. 13, no. 9, p. 5572, 2023.

[34] X. Wang, Y. Wang, and D. Zhang, "Complex emotion recognition via facial expressions with label noises self-cure relation networks," *Computational Intelligence and Neuroscience*, vol. 2023, Article ID 7850140, 10 pages, 2023.

[35] D. Zhu, Y. Fu, X. Zhao, X. Wang, and H. Yi, "Facial emotion recognition using a novel fusion of convolutional neural network and local binary pattern in crime investigation," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 2249417, 14 pages, 2022.

[36] N. Farhoumandi, S. Mollaey, S. Heysieattalab, M. Zarean, and R. Eyvazpour, "Facial emotion recognition predicts alexithymia using machine learning," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 2053795, 10 pages, 2021.

[37] Kaggle, "CKPLUS: CK+ dataset for facial expression recognition," 2018, https://www.kaggle.com/datasets/shawon10/ckplus.

[38] Kaggle, "FER-2013 dataset for facial expression recognition," 2018, https://www.kaggle.com/datasets/msambare/fer2013.

[39] M. Lyons, M. Kamachi, and J. Gyoba, "The Japanese female facial expression (JAFFE) dataset," 1998, https://zenodo.org/record/3451524.

[40] D. O. Oyewola, E. G. Dada, S. Misra, and R. Damaševičius, "Detecting cassava mosaic disease using a deep residual convolutional neural network with distinct block processing," *PeerJ Computer Science*, vol. 7, 2021.

[41] D. O. Oyewola, E. G. Dada, S. Misra, and R. Damaševičius, "A novel data augmentation convolutional neural network for detecting malaria parasite in blood smear images," *Applied Artificial Intelligence*, vol. 36, no. 1, Article ID 2033473, 2022.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: transformers for image recognition at scale," 2020, https://arxiv.org/abs/2010.11929.

[43] N. Dong, L. Zhao, C. H. Wu, and J. F. Chang, "Inception v3 based cervical cell classification combined with artificially extracted features," *Applied Soft Computing*, vol. 93, Article ID 106311, 2020.

[44] M. Mateen, J. Wen, N. Nasrullah, S. Song, and Z. Huang, "Fundus image classification using VGG-19 architecture with PCA and SVD," *Symmetry*, vol. 11, no. 1, p. 1, 2018.

[45] S. Han, Z. Meng, A. S. Khan, and Y. Tong, "Incremental boosting convolutional neural network for facial action unit recognition," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[46] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 558–565, IEEE, Washington, DC, USA, May 2017.

[47] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in

Proceedings of the IEEE international conference on computer vision, pp. 2983–2991, Santiago, Chile, December 2015.

[48] G. Meena, K. K. Mohbey, A. Indian, and S. Kumar, "Sentiment analysis from images using vgg19 based transfer learning approach," *Procedia Computer Science*, vol. 204, pp. 411–418, 2022.

[49] M. I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.

[50] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, "Deep learning approaches for facial emotion recognition: a case study on FER-2013," *Advances in Hybridization of Intelligent Methods: Models, Systems and Applications*, Springer, Berlin, Germany, pp. 1–16, 2018.

[51] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: aff-wild2, multi-task learning and arcface," 2019, https://arxiv.org/abs/1910.04855.

[52] Y. Shima and Y. Omori, "Image augmentation for classifying facial expression images by using deep neural network pretrained with object image database," in *Proceedings of the 3rd International Conference on Robotics, Control and Automation*, pp. 140–146, Chengdu, China, August 2018.

[53] H. Wang, S. Wei, and B. Fang, "Facial expression recognition using iterative fusion of MO-HOG and deep features," *The Journal of Supercomputing*, vol. 76, no. 5, pp. 3211–3221, 2020.

[54] B. Niu, Z. Gao, and B. Guo, "Facial expression recognition with LBP and ORB features," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 8828245, 10 pages, 2021.