Facial Expression Recognition using deep Convolutional Neural Network

Hiwa Wahab Ahmed[1], Asim Majeed Murshid[2]

[1, 2] Department of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq
Email: stcm22014@uokirkuk.edu.iq[1], dr.asim.majeed@uokirkuk.edu.iq[2]

Received: 15.07.2024 Revised: 20.08.2024 Accepted: 15.09.2024

ABSTRACT
The system that reads facial emotions from images is called Facial Expression Recognition (FER) or Facial Recognition (FR). This task provides insight into an individual's emotional condition. Deep learning-based FER is a field of AI approaches that is developing quickly. the goal of facial expression analysis is to recognize human emotions. This method automatically recognizes and extracts discriminative elements from face images using deep neural networks, namely CNNs. FER systems are now far more accurate and resilient thanks to deep learning models. Large-scale annotated datasets are used to train these models. Seven pre-trained deep learning models—InceptionResNetV2, InceptionV3, MobileNetV2, ResNet101, ResNet50, Xception, and GoogleNet—were used for the FER system in this thesis. Each utilized model varies based on modified parameters. We used both the Adam and SGD optimizers to evaluate each model based on a range of batch sizes, including 16, 32, 64, and 128. The FER-2013 large dataset, which consists of small-sized grayscale images, was used to assess the models. For the purpose of feeding the deeply pre-trained deep learning models, we transform the dataset to RGB color space. We obtained 65.47%, 65.89%, 64.25%, and 65.05% for metrics accuracy, precision, recall, and f1-score, respectively, in the InceptionResNetV2 model using the Adam optimizer and batch size 128. When compared to the most advanced model, we may improve the model's correctness in terms of performance.

Keywords:FER, CNN, Face Detection, deep learning.

1. INTRODUCTION

Facial expressions are the common way that all people show emotion. Numerous attempts have been made to develop automated techniques for studying facial expressions [1, 2]. Ekman et al. [3] identified seven basic emotions century (fear, terrified, happy, sad, scorn, disgust, and surprise from FER2013 dataset) [4], every human being, regardless of culture, has been evolving seven fundamental emotions [5]. FER has changed a lot because of AI, especially deep learning and machine learning. These technologies make it possible to automate the study of facial emotions, which increases accuracy and opens up new uses. AI, especially machine learning and deep learning, has changed the way we recognize face expressions. AI can improve efficiency, scale, and the number of ways it can be used. Deep learning has pushed the limits of what is possible with machine learning. In addition, deep learning can make FER systems more complicated and effective. Dealing with the problems that come up with big data needs involves balancing classes and computing needs. There are two main AI-based FER method approaches. The first strategy uses manually created feature vectors to do recognition [6, 7]. On the other hand, the other strategy is an end-to-end technique that leverages deep neural network-based automatically derived characteristics to achieve recognition [8-10]. In order to address and resolve, the following are our primary contributions:

1. Examine the FER datasets that are currently available, and use the most recent one to test and refine our suggested method.
2. To assess and train FER, choose a large dataset.
3. Adjust the fine-tuning settings to improve the accuracy of the model's performance.

2. Related Work

Several researchers are using deep learning models, machine learning algorithms, or a mix of deep learning and machine learning approaches for FER systems. They used a variety of datasets and models to get impressive results.

Due to deep learning's strong automated identification capability, researchers have focused on it during the last ten years, despite the remarkable success of conventional face recognition techniques including

classical feature extraction and machine learning algorithms [11-16]. We will discuss some current FER research in this area, showcasing deep learning techniques that have been suggested to improve detection [17].

Over the course of various investigations, machine learning techniques were applied for FER [18-23]. Chen and Jenkins (2017) introduced three distinct face recognition techniques that use machine learning algorithms. For feature extraction and dimension reduction, they used machine learning methods including SVM, KNN, and LDA based on PCA [18]. Comparisons and evaluations of recognition accuracy and running time show that PCA + SVM yields the best recognition result, which is over 95%, for certain training data and eigenface sizes. Furthermore, PCA + KNN balances recognition accuracy and speed of operation [21].A summary of the recent literature reviews on FER systems is shown in Table 1, with an emphasis on suggested solutions, dataset types, and outcomes assessed by model performance accuracy.

Table 1. An overview of the literature on FER

Ref.	Proposed Solutions	Dataset	Obtained Results (Accuracy)
[21]	SVM+PCA KNN +PCA LDA+PCA	ORL	SVM+PCA = 95%
[20]	A combination of global features (PCA, DWT), local features (LBP, HOG, and Gabor wavelet), with Classifiers (NN, SVM)	Essex	99.45%
[24]	CNN	MultiPie	94.7%
[25]	CNN	CK+, JAFFE, BU-3DFE	CK+ = 96.76%
[26]	CNN with preprocessing	FIVE FACIAL EXPRESSION IMAGE	88%
[27]	CNN+LSTM	CK+	99.43%
[28]	(AlexNet and VGG-16) + SVM	MLF-W-FER	63.4% 65.5%
[29]	Resnet18	RAF	97.75%
[30]	CNN, BoVW with KNN, SVM	FER 2013, FER+, AffectNet	87.76%
[31]	CNN	RaFD	94.44%
[32]	CNN	JAFFE	99.66%
[33]	CNN with LBP	CK+ Oulu-CASIA JAFFE NCUFE	94.63% 98.52% 94.33% 98.68%
[34]	Individual CNN, Global CNN	EMOTIW	83.9% 80.9%
[35]	CNN	JAFFE	98.65%
[36]	CNN with LSTM	Oulu-CASIA	88.3%

3. METHODOLOGY

3.1 Dataset

The FER2013 dataset is generated by using expression-related keywords in a Google picture search. Each picture in the collection is composed of a fixed 48 by 48-pixel grayscale image. Every picture has a face that is about in the middle and occupies the same amount of space since the faces are automatically captured. Depending on the emotion shown in the facial expression, each face must be placed into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, and 6=Neutral). 35,886 images of people in various states of feeling are available. The dataset is unbalanced; there are 3,589 cases in the public test set and 28,709 instances in the training set [37]. This paper makes both balanced (undersampling [38]) and unbalanced use of the FER2013 dataset. Figure 1 shows a selection of the photos from the FER2013 dataset.

Figure 1. Some randomly selected images from the FER2013 dataset[37].

3.2 Proposed Structure

In this paper, we use the FER-2013 dataset to train and assess seven pre-trained deep learning models, namely InceptionResNetV2, InceptionV3, MobileNetV2, ResNet101, ResNet50, Xception, and GoogleNet, for FER.

The integrated deep learning model technique is shown in Figure 3-3. The goal of this approach is to identify the FER test pictures as accurately, quickly, and with the fewest possible mistakes. In order to identify the FER, our suggested model consists of two steps: first, preprocessing the grayscale FER-2013 dataset into RGB (red, green, and blue) format; second, feed all RGB FER-2013 dataset to seven deep learning models, such as InceptionResNetV2, InceptionV3, MobileNetV2, ResNet101, ResNet50, Xception, and GoogleNet.

The models developed in this study are built using the same procedure that every CNN-based model does. The data—that is, the values of the RGB photographs—has been input into the model after the loading of the images. The feature extraction and classification processes are carried out by the model itself. Ultimately, the model has anticipated the projected production, as Figure 2 illustrates.

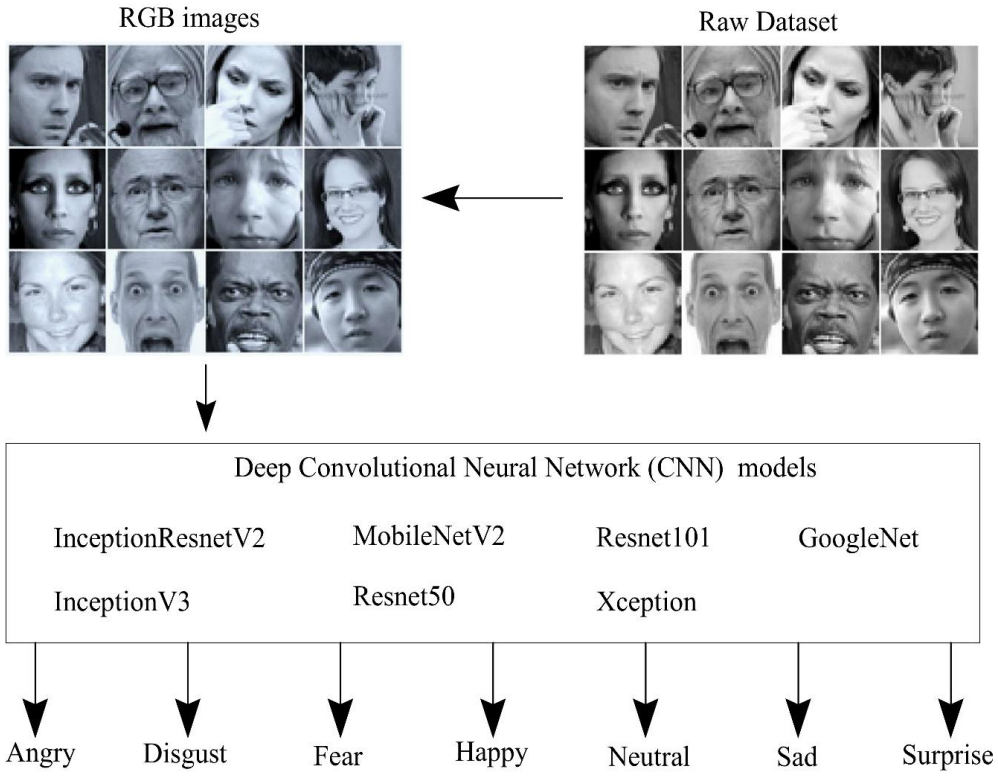

Figure 2. The proposed structure

3.2.1 Inception Res NetV2

A deep CNN architecture called InceptionResNetV2 [39] combines residual connections with the Inception design. The advantages of residual connections, which help train extremely deep networks, and initiation modules, which record multi-scale data, are combined in this hybrid model.

A deep CNN architecture called InceptionResNetV2 combines residual connections with the Inception design. The advantages of residual connections, which help train extremely deep networks, and initiation modules, which record multi-scale data, are combined in this hybrid model. InceptionResNetV2 is a potent model that can handle a variety of challenging picture identification tasks because it leverages the architectural breakthroughs of both Inception and ResNet.

3.2.2 InceptionV3

As a member of Google's Inception network family, inceptionV3 [40] is a deep convolutional neural network architecture. It has a reputation for doing well on tasks involving picture recognition. Numerous enhancements and adjustments have been made to InceptionV3 to increase its precision and effectiveness. The following are some of InceptionV3's key components.

1. InceptionV3 makes use of Inception modules, which concatenate the outputs after applying many convolutional filters of varying sizes to the same input. As a result, the network can record characteristics at different sizes.
2. Factored convolutions are used by InceptionV3 to save computational costs and increase efficiency. For example, a big convolutional filter may be factorized into smaller filters.
3. Auxiliary classifiers are used into InceptionV3 to enhance convergence during training. These are simply extra network branches that act as regularizes and aid in the propagation of gradients.
4. Batch normalization is widely used across the network to increase training stability and speed.
5. To avoid overfitting, InceptionV3 uses label smoothing during training. This method may enhance generalization by altering the ground truth labels to reduce the model's confidence.

3.2.3 MobileNetV2

MobileNetV2 [41], a lightweight CNN architecture, was created especially for embedded and mobile vision applications. It was created by Google researchers. This model is great for devices with limited resources because it strikes a good mix between model size and accuracy.

A number of important parts of MobileNetV2 make picture recognition more efficient. All of these things are necessary to keep the model's accuracy high while keeping its computing requirements low.

3.2.4 ResNet50 and ResNet101

The ResNet design is typically divided into four sections, each containing numerous residual blocks at different depths [42]. In order to reduce the spatial dimensions of the input, the network's initial phase includes a single convolutional layer, which is then followed by max pooling. In the second half of the network, there are a total of 64 filters. Moving on to the third and fourth sections, we have 128 and 256 filters, respectively. The final section of the Network consists of global average pooling and a fully connected layer that generates the output.

There are other variations of the ResNet design, such as ResNet-18, ResNet-34, ResNet-50 (see Figure 3-13), ResNet-101, and ResNet-152. The number in each version is the same as the number of Network layers. ResNet-101, for instance, has 101 layers, while ResNet-50 has 50. We used ResNet-50 and ResNet-101 in this paper.

3.2.5 Xception

"Extreme Inception," or "Xception," [43] is an acronym signifying a significant turning point in CNN design. 2017 saw the release of Xception, an advancement of the Inception architecture as shown in Figure 3-14.

Xception is distinguished by its depthwise separable convolutions. Depthwise separable convolutions separate these processes, in contrast to typical convolutions, which act on both spatial and depth dimensions concurrently. This preserves representational capacity while drastically cutting down on the number of parameters and computing expense.

A collection of separable convolutional blocks forms the foundation of Xception's design. A linear transformation via a pointwise convolution is the first step in each block, which is a depthwise separable convolution. The model's capacity to capture intricate patterns with fewer parameters is improved by its modular architecture.

3.2.6 GoogleNet

According to [44], GoogLeNet is a deep CNN. Specifically, the architecture of this model was created in a way that permits network expansion in both depth and breadth while maintaining computational capacity. Nine Inception blocks make up the total of 22 layers in the VGG model. There are four parallel routes in each Inception block where convolution layers with various kernel sizes are applied.

The first route makes use of a convolutional layer with a 1×1 window size. A 1×1 convolutional layer is used in the second and third routes, followed by two costly 3×3 and 5×5 convolutions. The model complexity is decreased by lowering the number of filter channels with the use of the 1×1 convolution.

3.3 Evaluation Criteria

Metrics like accuracy, precision, recall, and f1-score are crucial for assessing a deep learning model. Through the emphasis on its strengths and weaknesses, this assessment contributes to the model's improvement.

3.3.1 Accuracy

When assessing the classification models, which predict the label, accuracy is a key factor (Equation 1). Use of accuracy as a statistic for a classification model requires a balanced dataset. This figure determines how well the model predicts. Computed using the ratio of true positives (TP) to true negatives (TN) to the total number of samples.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \quad (1)$$

3.3.2 Precision

The ratio of all of the model's positive forecasts to the proportion of true positive forecasts yields the model's accuracy. The computation of it is the TP to FP total ratio. Division of the total number of true positives by the number of false positives and true positives is the precision formula (Equation 2).

$$\text{precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False positive (FP)}} \quad (2)$$

3.3.3 Recall

Recall measures the proportion of accurate positive predictions compared to all actual positive instances. It is also known as sensitivity or the true positive rate. It is calculated by dividing the number of true positives by the sum of false negatives [45].Recall (Equation 3) is determined by calculating the ratio of true positives to the sum of true positives and false negatives.

$$Recall = \frac{\text{True Positive (TP)}}{\text{True positive (TP)} + \text{False Negative (FN)}} \quad (3)$$

3.3.4 F1-score

F1-score is an alternative machine learning evaluation statistic that assesses the predictive ability of a model by concentrating on its performance within each class as opposed to its overall performance, as accuracy does. Combining the precision and recall scores of a model into a single metric, the F1-score (equation 4), has led to its widespread application in recent research [46]. The accuracy statistic quantifies the frequency with which a model accurately predicts the entire dataset.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

While precision and recall concentrate on the caliber of positive and negative predictions, respectively, accuracy assesses the overall correctness of the model's predictions. Because it strikes a compromise between recall and accuracy, the F1 Score is a more thorough statistic for assessing classification models.

The InceptionResNetV2, InceptionV3, MobileNetV2, ResNet101, ResNet50, Xception, and GoogleNet models were implemented using the Python programming language to transform grayscale images to RGB images. 64 GB of RAM, an Intel Core i7-8700K 3.7GHz CPU, and two NVIDIA GeForce GTX 1080 Ti graphics processing units operating in parallel were used for all training techniques.

4. Experimental Results and Discussion

This section discusses and analyzes the outcomes of our deep learning classifier-based suggested strategy. The FER-2013 dataset, which is an unbalanced dataset with tiny images with dimensions of 48 by 48, was used, to train and evaluate the suggested model. The subset of the dataset was separated using cross-validation and percentage split to create training and testing sets. The dataset was divided into two sets, of which 80% were used for training and 20% for testing. We evaluated our models using the seven pretrained deep learning models, including InceptionResNetV2, InceptionV3, MobileNetV2, ResNet101, ResNet50, Xception, and GoogleNet that used ImageNet weight.

4.1 Results

There are two scenarios in the suggested technique. In the first, we used the imbalanced FER-2013 dataset to train and assess all specified models, meaning we didn't change the number of images in each class. There are 35887 images in the FER-2013 data collection.

In the second case, we used undersampling method [38] to initially balance the FER-2013 dataset. The number of images in each class decreases to the lowest class in the dataset while using undersampling method. The class "disgust" in the FER-2013 dataset contains the fewest images—547 total. It reduces all classes to 547 images. With balancing data, FER-2013 has a total of 3829 images.

4.2 Discussion

Based on seven deep learning models that have been trained beforehand, Table 2 presents the acquired results with data that is imbalanced. The best outcomes for each model are shown in each row. After selecting the top result and noting it in Table 2, we received 8 results in each row, demonstrating the employment of both optimizers, Adam [47]and SGD [48], with different batch sizes of 16, 32, 64, and 128. Each class has a variable number of images, and the dataset is imbalanced. Some classes have nine times as many images as other classes, which leads to significant imbalances and impacts how the models are trained and evaluated. Because of this, we used an unbalanced dataset to train and test all of the models, and although all of the models produced results over 50% accuracy, the InceptionResNetV2 model was able to reach a high model performance accuracy of 65.47% accuracy.Figure 3 presents a comparative analysis of several models using imbalanced data.The ROC curve was utilized to identify models with the highest accuracy when dealing with imbalanced data (see figures 4).

Table 2. The obtained results for various deeplearning models based on imbalanced data

Models	Epochs	Optimizer	Learning rate	Batch Size	Number of Iterations	image size	Accuracy	Precision	Recall	F1-score
InceptionResNetV2	700	Adam	0.001	128	225	75*75*3	65.47	65.89	64.25	65.05
InceptionV3	700	Adam	0.001	128	225	75*75*3	61.74	61.33	61	61.16
MobileNetV2	700	Adam	0.001	32	898	48*48*3	59.24	59	58.14	58.56
ResNet50	700	SGD	0.01	16	1795	48*48*3	54.58	54.39	54.67	54.52
ResNet101	700	SGD	0.01	64	449	48*48*3	63.01	63.47	63.08	63.27
Xception	700	SGD	0.01	16	1795	48*48*3	60.24	61.32	61.14	61.22
GoogleNet	700	SGD	0.01	32	898	48*48*3	50.77	50.16	50.5	50.32

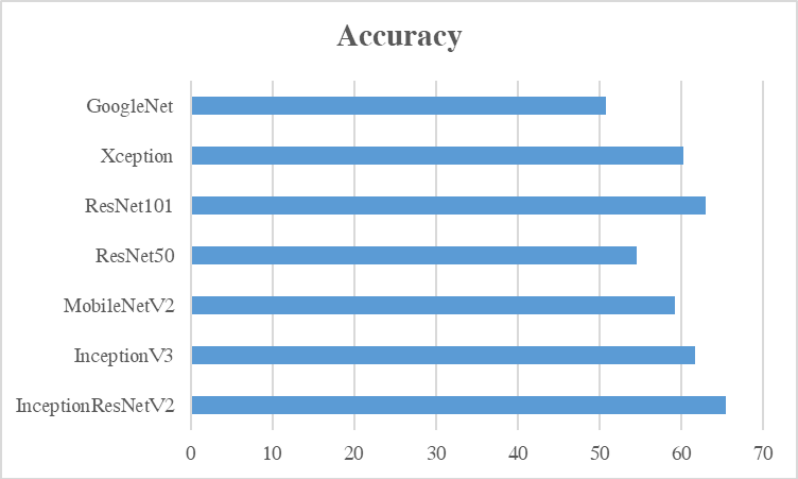

Figure 3. Comparison different models using imbalanced data

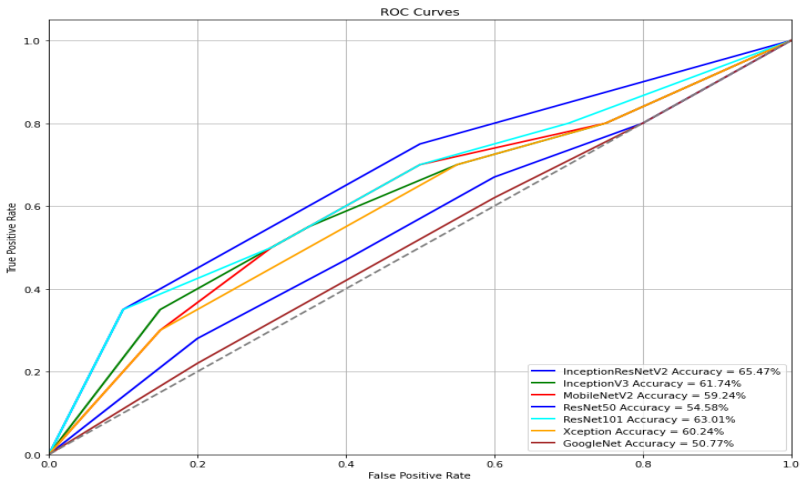

Figure 4. ROC curve for utilized models with the highest accuracy using imbalanced data

We used the undersampling approach on balanced data, which yielded the findings shown in Table 3. Each model was trained and tested using various batch sizes and optimizers. Table 3 lists the top outcomes for each model along with additional details.Figure 5 presents a comparative analysis of several models using balanced data.The ROC curve was utilized to identify models with the highest accuracy when dealing with balanced data (see figures 6).

Table 3. The obtained results for various deeplearning models based on balanced data

Models	Epochs	Optimizer	Learning rate	Batch Size	Number of Iterations	image size	Accuracy	Precision	Recall	F1-score
InceptionResNetV2	500	SGD	0.01	32	120	75*75*3	68.49	69.47	69.13	69.29
InceptionV3	500	Adam	0.001	64	60	75*75*3	64.71	64.8	64.03	64.41
MobileNetV2	500	Adam	0.001	64	60	48*48*3	64.89	63.12	63.02	63.07
ResNet50	500	Adam	0.001	64	60	48*48*3	58.79	59.13	59.62	59.37
ResNet101	500	SGD	0.01	64	60	48*48*3	65.95	65.9	65.91	65.9
Xception	500	SGD	0.01	128	30	48*48*3	63.21	63.74	63.56	63.64
GoogleNet	500	SGD	0.01	128	30	48*48*3	53.91	53.28	53.67	53.47

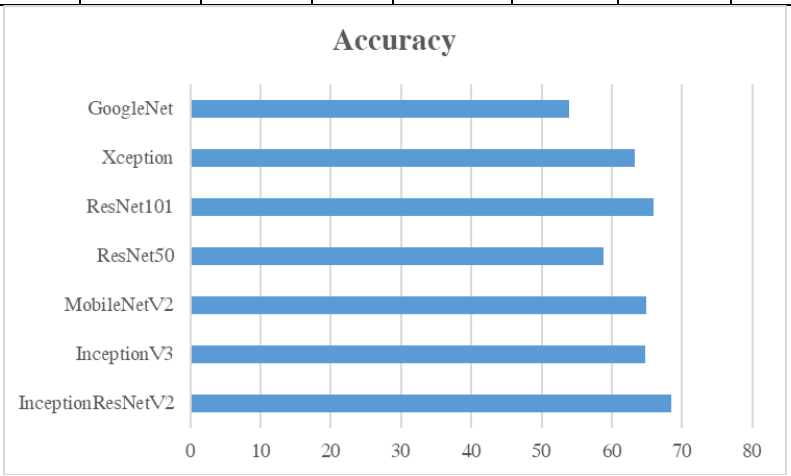

Figure 5. Comparison different models using balanced data

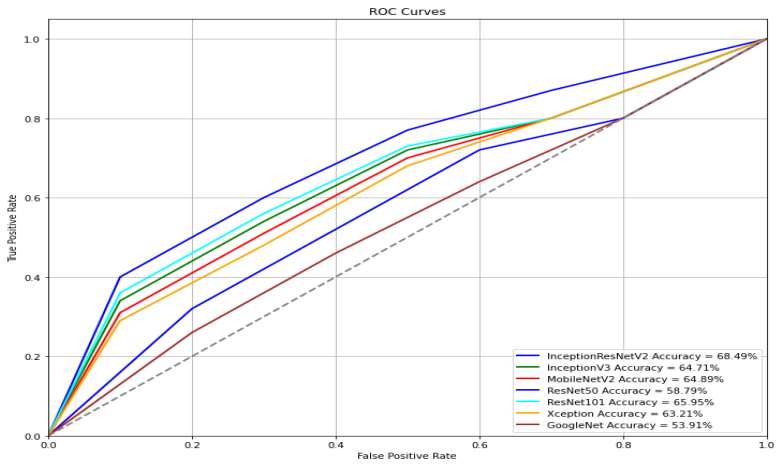

Figure 6. ROC curve for utilized models with the highest accuracy using balanced data

In this investigation as shown in Table 4, the performance of an existing model [4] for face image classification on unbalanced data is compared to that of the InceptionResNetV2 model employing several optimizers. Metrics for recall, accuracy, precision, and F1-score are shown in the table. InceptionResNetV2 performs better than the current model [4] in all measures for both setups, indicating improved performance for face image classification on unbalanced data. When compared to the SGD optimizer, the Adam optimizer performs the best, demonstrating its ability to handle unbalanced datasets. The InceptionResNetV2 model outperforms the existing model [4]in all measures, including accuracy, precision, recall, and F1-score. This difference is rather noticeable. The Adam optimizer is the optimal choice in the suggested task since it yields the best outcomes. This comparison shows how resilient InceptionResNetV2 is and how important optimizer selection is for achieving optimal performance on imbalanced data. We use unbalanced data to compare our suggested models with the most advanced ones, as seen in Figure 7.The confusion matrix for the InceptionResnetV2 model, which achieved the greatest accuracy utilizing unbalanced data, is shown in Figure 8. It provides detailed information about the evaluation values for each class.

Table 4. Our suggested model's comparison with the current facial image classification based on imbalanced data

Models	Epochs	Optimizer	Learning rate	Batch Size	Number of Iterations	image size	Accuracy	Precision	Recall	F1-score
[4]							64.3	62.2	59.4	59.6
InceptionResNetV2	700	Adam	0.001	128	225	75*75*3	**65.47**	**65.89**	**64.25**	**65.05**
InceptionResNetV2	700	SGD	0.01	128	225	75*75*3	**64.98**	**63.87**	**63.54**	**63.7**

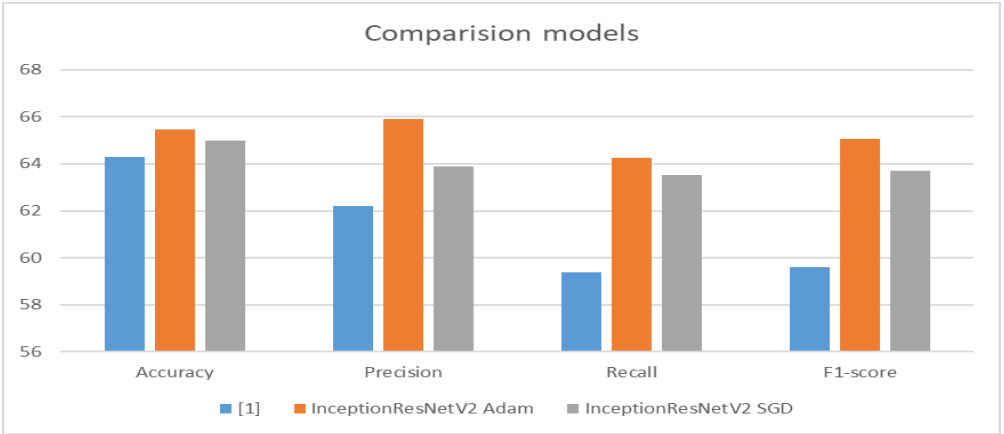

Figure 7. Comparison of our proposed models with state-of-the-art models using imbalanced data

TARGET / OUTPUT	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	703 9.79%	82 1.14%	54 0.75%	62 0.86%	2 0.03%	41 0.57%	14 0.20%
Disgust	23 0.32%	54 0.75%	1 0.01%	1 0.01%	8 0.11%	13 0.18%	11 0.15%
Fear	45 0.63%	37 0.52%	797 11.10%	76 1.06%	1 0.01%	0 0.00%	68 0.95%
Happy	98 1.37%	76 1.06%	31 0.43%	1227 17.09%	138 1.92%	193 2.69%	11 0.15%
Neutral	80 1.11%	119 1.66%	36 0.50%	127 1.77%	784 10.92%	74 1.03%	13 0.18%
Sad	231 3.22%	87 1.21%	60 0.84%	171 2.38%	61 0.85%	605 8.43%	32 0.45%
Surprise	56 0.78%	27 0.38%	101 1.41%	19 0.26%	56 0.78%	43 0.60%	529 7.37%

Figure 8. Confusion matrix of the inceptionresnetv2 using imbalanced data

CONCLUSION

A challenging task in computer vision is FER. FER models may be used to classify a broad spectrum of human emotions. In this work, we assessed several deep pre-trained models on both balanced and unbalanced observations from the FER-2013 dataset. Xception, GoogleNet, ResNet50, ResNet101, InceptionResNetV2, and MobileNetV2 were some of these models. The FER-2013 dataset was used to assess the models; it was a large dataset that had data that was not balanced. There are some low-quality images in the collection. The results indicate that increasing the model's performance accuracy is mostly dependent on parameter optimization. Our results clarify the usefulness of these models and demonstrate the influence of dataset balance on classification performance. The main goal is to test and train datasets using various batch sizes (16, 32, 64, and 128). The goal of these batch sizes is to improve the accuracy of the model's performance. The results of using the Adam and SGD optimizers indicated that improving model performance could be a solution. Depending on the outcomes, deep learning settings may need to be adjusted to improve model accuracy. In the future, the accuracy of the model performance may be improved by combining machine learning and deep learning models into one merged model.

REFERENCE

[1] A. Kartali, M. Roglić, M. Barjaktarović, M. Đurić-Jovičić, and M. M. Janković, "Real-time algorithms for facial emotion recognition: a comparison of different approaches," in 2018 14th Symposium on Neural Networks and Applications (NEUREL), 2018: IEEE ,pp. 1-4 .

[2] A. P. Lim, G. P. Kusuma, and A. Zahra, "Facial emotion recognition using computer vision," in 2018 Indonesian Association for Pattern Recognition International Conference (INAPR), 2018: IEEE, pp. 46-50 .

[3] R. S. Deshmukh, V. Jagtap, andS. Paygude, "Facial emotion recognition system through machine learning approach," in 2017 international conference on intelligent computing and control systems (iciccs), 2017: IEEE, pp. 272-277 .

[4] B. Kaur, P. K. Jindal, R. Gill, P. Sharma, and M. Kaur" ,Facial Recognition System using Convolutional Neural Networks (CNN)," in 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), 2023: IEEE, pp. 1-6 .

[5] K .Sarvakar, R. Senkamalavalli, S. Raghavendra, J. S. Kumar, R. Manjunath, and S. Jaiswal, "Facial emotion recognition using convolutional neural networks," Materials Today: Proceedings, vol. 80, pp. 3560-3564, 2023.

[6] Y. Liu, Y. Li, X. Ma, and R. Song, "Facial expression recognition with fusion features extracted from salient facial areas," Sensors, vol. 17, no. 4, p. 712, 2017.

[7] M. Kopaczka, R. Kolk, J. Schock, F. Burkhard, and D. Merhof, "A thermal infrared face database with facial landmarks and emotion labels," IEEE Transactions on Instrumentation and Measurement, vol. 68, no. 5, pp. 1389-1401, 2018.

[8] S. Vignesh, M. Savithadevi, M. Sridevi, and R. Sridhar, "A novel facial emotion recognition model using segmentation VGG-19 architecture," International Journal of Information Technology, vol. 15, no. 4, pp. 1777-1787, 2023.

[9] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," in 2020 25Th international conference on pattern recognition (ICPR), 2021 :IEEE, pp. 4513-4519 .

[10] S. K. KM, R. Rajendran, Q. Wan, K. Panetta, and S. S. Agaian, "TERNet: A deep learning approach for thermal face emotion recognition," in Mobile Multimedia/Image Processing, Security, and Applications 2019, 2019, vol. 10993: SPIE, pp. 45-51 .

[11] E. M. Onyema, P. K. Shukla, S. Dalal, M. N. Mathur, M. Zakariah, and B. Tiwari, "Enhancement of patient facial recognition through deep learning algorithm: ConvNet," Journal of Healthcare Engineering, vol. 2021, no. 1, p. 5196000, 20.21

[12] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, "Facial expression recognition based on deep learning," Computer Methods and Programs in Biomedicine, vol. 215, p. 106621, 2022.

[13] S. M. S. Abdullah and A. M. Abdulazeez, "Facial expression recognition based on deep learning convolution neural network: A review," Journal of Soft Computing and Data Mining, vol. 2, no. 1, pp. 53-65, 2021.

[14] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," Neural Computing and Applications, vol. 35, no. 32, pp. 23311-23328, 2023.

[15] A. Khopkar and A. Adholiya, "Facial expression recognition using CNN with Keras," Bioscience Biotechnology Research Communications, vol. 14, no ,5 .pp. 47-50, 2021.

[16] J. D. Bodapati, U. Srilakshmi, and N. Veeranjaneyulu, "FERNet: a deep CNN architecture for facial expression recognition in the wild," Journal of The institution of engineers (India): series B, vol. 103, no. 2, pp. 439-448, 2022.

[17] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," Procedia Computer Science, vol. 175, pp. 689-694, 2020.

[18] S. Sharma, M. Bhatt, and P. Sharma, "Face recognition system using machine learning algorithm," in 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020: IEEE, pp. 1162-1168 .

[19] M. Atay, H. Gipson, T. Gwyn, and K. Roy, "Evaluation of gender bias in facial recognition with traditional machine learning algorithms," in 2021 IEEE Symposium Series on Computational Intelligence (SSCI), 2021: IEEE, pp. 1-7 .

[20] K. Raju, B. Chinna Rao, K. Saikumar, and N. Lakshman Pratap, "An optimal hybrid solution to local and global facial recognition through machine learning," A fusion of artificial intelligence and internet of things for emerging cyber systems, pp. 203-226, 2022.

[21] J. Chen and W. K. Jenkins, "Facial recognition with PCA and machine learning methods," in 2017 IEEE 60th international Midwest symposium oncircuits and systems (MWSCAS), 2017: IEEE, pp. 973-976 .

[22] Y. K. Bhatti, A. Jamil, N. Nida, M. H. Yousaf, S. Viriri, and S. A. Velastin, "Facial expression recognition of instructor using deep features and extreme learning machine," Computational Intelligence and Neuroscience, vol. 2021, no. 1, p. 5570870, 2021.

[23] Z. Song, "Facial expression emotion recognition model integrating philosophy and machine learning theory," Frontiers in Psychology, vol. 12, p. 759485, 2021.

[24] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in 2016 IEEE Winter conference on applications of computer vision (WACV), 2016: IEEE, pp. 1-10 .

[25] A. T. Lopes, E. De Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," Pattern recognition, vol. 61, pp. 610-628, 2017.

[26] Y. Luo, J. Wu, Z. Zhang, H. Zhao, and Z. Shu, "Design of Facial Expression Recognition Algorithm Based on CNN Model," in 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA), 2023: IEEE, pp. 580-583 .
[27] R. Febrian, B. M. Halim, M. Christina, D. Ramdhan, and A. Chowanda, "Facialexpression recognition using bidirectional LSTM-CNN," Procedia Comput. Sci, vol. 216, no. 2022, pp. 39-47, 2023.
[28] H. 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.
[29] D.-H. Lee and J.-H. Yoo, "CNN Learning Strategy for Recognizing Facial Expressions," IEEE Access, 2023.
[30] 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.
[31] G. Yolcu et al., "Facial expression recognition for monitoring neurological disorders based on convolutional neural network," Multimedia Tools and Applications, vol. 78, pp. 31581-31603, 2019.
[32] Y. Wang, Y. Li, Y. Song, and X. Rong, "The influence of the activation function in a convolution neural network model of facial expression recognition," Applied Sciences, vol. 10, no. 5, p. 1897, 2020.
[33] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based CNN for facial expression recognition," Neurocomputing, vol. 411, pp. 340-350, 2020.
[34] L. Tan, K. Zhang, K. Wang, X. Zeng, X. Peng, and Y. Qiao, "Group emotion recognition with individual facial emotion CNNs and global image based CNNs," in Proceedings of the 19th ACM international conference on multimodal interaction, 2017, pp. 549-552 .
[35] A. Jaiswal, A. K. Raju, and S. Deb, "Facial emotion detection using deep learning," in 2020 international conference for emerging technology (INCET), 2020: IEEE, pp. 1-5 .
[36] F. An and Z. Liu, "Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM," The visual computer, vol. 36, no. 3, pp. 483-498, 2020.
[37] K. Liu, M. Zhang, and Z. Pan, "Facial expression recognition with CNN ensemble," in 2016 international conference on cyberworlds (CW), 2016: IEEE, pp. 163-166 .
[38] H. He and Y. Ma, "Imbalanced learning: foundations, algorithms, and applications," 201.3
[39] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Proceedings of the AAAI conference on artificial intelligence, 2017, vol. 31, no. 1 .
[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818-2826 .
[41] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision andpattern recognition, 2016, pp. 770-778 .
[43] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251-1258 .
[44] C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1-9 .
[45] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PloS one, vol. 10, no. 3, p. e0118432, 2015.
[46] B.-C. Yan, H.-W. Wang, S.-W. F. Jiang, F.-A. Chao, and B. Chen, "Maximum f1-score training for end-to-end mispronunciation detection and diagnosis of L2 English speech," in 2 022IEEE International Conference on Multimedia and Expo (ICME), 2022: IEEE, pp. 1-5 .
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[48] G. Montavon, G. Orr, and K.-R. Müller, Neural networks: tricks of the trade. springer, 2012.