FACIAL EMOTION RECOGNITION IN IMBALANCED DATASETS

Sarvenaz Ghafourian, Ramin Sharifi and Amirali Baniasadi

Department of Electrical and Computer Engineering,
University of Victoria, Victoria, Canada

ABSTRACT

The wide usage of computer vision has become popular in the recent years. One of the areas of computer vision that has been studied is facial emotion recognition, which plays a crucial role in the interpersonal communication. This paper tackles the problem of intraclass variances in the face images of emotion recognition datasets. We test the system on augmented datasets including CK+, EMOTIC, and KDEF dataset samples. After modifying our dataset, using SMOTETomek approach, we observe improvement over the default method.

KEYWORDS

Emotion Recognition, Residual Network, VGG.

1. INTRODUCTION

Since the human face plays an integral part in expressing a person's mental state, facial expression analysis is a significant research focus with numerous potential uses. Scientists from many areas like psychology, finance, marketing, and engineering have been greatly interested in this subject due to the practical benefits.

Artificial intelligence is becoming more prevalent in many aspects of human life. The technologies are adapted to the needs of human beings, and artificial intelligence is what makes this adaptation between technology and humans possible. While it may come easy for most humans to process emotions without any extra effort, computers have struggled with the idea of recognizing them automatically for decades. This challenge is due to face appearance changes caused by pose variations, illumination variations, camera quality, and angle changes. Research from different disciplines such as computer vision and machine learning focus on utilizing computers to categorize emotions exhibited by humans properly. In this work we focus on analyzing facial expressions. Specifically, we study the task of facial emotion recognition based on two deep learning models using our suggested dataset. We use various face images for seven emotions and improve the efficiency of emotion detection.

The face is defined as the front portion of the human head, from above by the scalp border, below by the corners and bottom edge of the lower jaw, and on the sides by the margins of the lower jaw branches and the base of the auricles [1].

Facial emotion recognition (FER) is a method to identify human expressions, which is one of the factors involved in emotion recognition. Emotions are inherent characteristics of people and play a significant part in social communication [2][3]. Humans show emotion in a variety of ways, including facial expressions [4], gestures, vocalizations, body language [5], and speech [6]. The

David C. Wyld et al. (Eds): AIAPP, NLPML, DMA, CRIS, SEC, CoSIT, SIGL - 2022
pp. 239-251, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120920

six basic emotions described by Eckman [7] are: happiness, sadness, fear, disgust, anger, and surprise.

Overfitting occurs when the Facial Emotion Recognition (FER) model is trained on imbalanced datasets, making the model less capable of performing FER tasks in real-world scenarios. As a result, overfitting due to lack of sufficient data remains a problem for most FER systems. Thus, we create an augmented dataset in this work to mitigate the overfitting problem and improving generalization.

Our goal is to identify an individual's emotion from observing their facial expressions. First, cropped headshots are extracted using the FaceNet architecture. Second, the extracted face images from three different datasets are used as a single dataset for the transfer-learning task on VGG-16 and ResNet-50.

In summary the contributions of this paper are:

- A comprehensive analysis of popular emotion recognition datasets, such as CK+, EMOTIC, and KDEF. We describe how images are categorized in each dataset.
- Creating a custom dataset consisting of the three above-mentioned datasets to cover a wide range of variations in face images. We explain how different images of our dataset are cropped to fit our criteria.
- Improving class imbalance problem in the custom emotion recognition dataset over VGG-16 and ResNet-50. We show how SMOTETomek technique improves the distinction accuracy over VGG-16 and ResNet-50 models.

This paper is organized as follows. Section 2 describes related works. Section 3 demonstrates the background. Section 4 represents experiments and results. Section 5 offers concluding remarks.

2. RELATED WORKS

The traditional approach to detecting emotions consists of a two-stage machine learning process. The first phase involves collecting characteristics from the pictures, and the second phase involves using a classifier, such as an SVM, neural network, or random forest, to determine the emotions.

The histogram of oriented gradients (HOG) [8], local binary patterns (LBP) [9], Gabor wavelets[10], and Haar features [11] are some of the prominent hand-crafted features utilized for face emotion identification. The appropriate emotion is then assigned to the image using a classifier.

While these methods work for small datasets, they start showing their limits when applied to more complex datasets, with higher intraclass variances. Moreover, there are some issues with face images when the face is partially visible [12].

The majority of contemporary computer vision research into recognizing people's emotional states is based on facial expression analysis. Psychologists, Ekman and Friesen, identified six fundamental emotions and multiple methods for recognizing them. The Facial Action Coding System is used in several of these approaches. Action Units (AU) are a collection of unique localized movements of the face that encode facial emotion. This approach uses a set of distinct localized facial movements known as Action Units to represent facial emotion [13].

Convolutional Neural Networks (CNNs) have been used in recent studies for emotion detection based on facial expression to recognize emotions and Action Units [14].

In response to the great success of deep learning and, in particular, CNNs for image classification and other vision challenges, a number of organisations have built deep learning-based facial expression recognition (FER) models [15]. Mollahosseini et al. showed that CNNs could recognize emotions accurately and achieve state-of-the-art results. The results are based on a zero-biased CNN on the expanded Cohn-Kanade dataset (CK+) and the Toronto Face Dataset (TFD). Mollahosseini also, suggested an FER neural network with two convolution layers, one max-pooling layer, and four inception layers, in each layer [16].

Aneja et al. in [17] created a model of facial expressions for stylized animated characters using deep learning. Their training included a network that represented human expressions, and a network that represented animated faces. The loopy network was first proposed by Liu in [18], noting the importance of feedback of the weak classifiers. Instead of using a strong classifier, a loop of weaker classifiers are used for emotion detection. They used their Boosted Deep Belief Network (BDBN) over CK+ and JAFFE datasets to achieve a higher accuracy.

In addition to determining the face characteristics, some studies [19] detect fundamental emotions using the position of shoulders. Schindler et al.[20] used a limited dataset of non-spontaneous postures obtained under controlled conditions to detect the six primary emotions.

Rather than identifying emotion categories, some more recent research on facial expression [21] employs the Valence, Arousal, Dominance (VAD) Emotional State Model continuous dimensions to describe emotions [22].

It should be noted that the majority of the past research is based on widely used facial expression recognition datasets, such as FER2013, the extended Cohn-Kanade (CK+), and the Japanese Female Facial Expression dataset (JAFFE). These datasets consist of frontal face images, and the photos lack any contextualized backgrounds and have fewer differences, such as spectacles or face masks. This makes the facial action units detection easier. However, we expect our model to perform on more challenging images as well. Images consisting of illumination, pose, occlusion, and low resolution ones are considered challenging images.

3. BACKGROUND

3.1. Face Recognition

The challenge of recognizing and validating people in an image by their faces is known as face recognition. Face recognition is sometimes defined as a four-step process that begins with face detection, then moves on to face alignment and feature extraction respectively, and ultimately face identification (Fig. 1) [23].

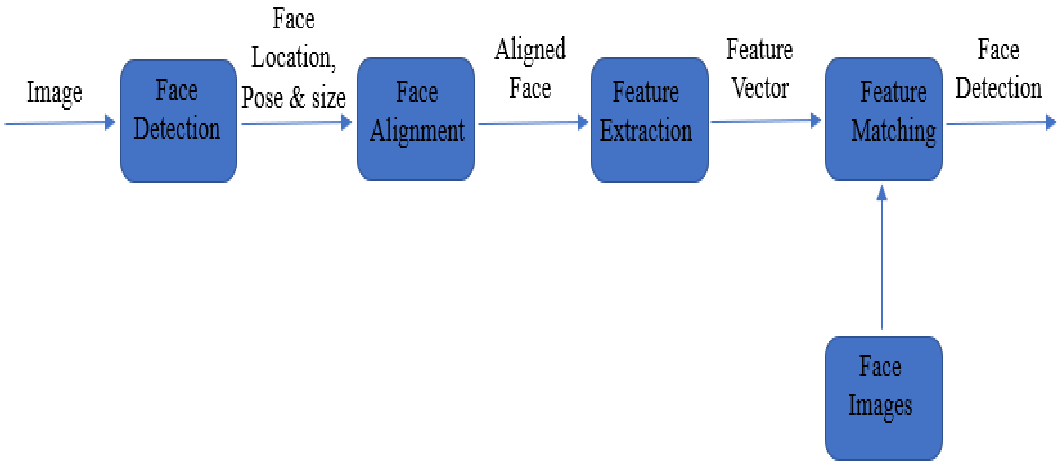


Figure 1. Face recognition processing flow [23].

Some of the face recognition task challenges include illumination that is caused by light fluctuation, pose which is the result of head movement and viewing angle changes, occlusion that is caused by blocking of one or more portions of a face, and low resolution.

3.2. Datasets

3.2.1. EMOTIC Dataset

Emotion in context (EMOTIC) [22] is a dataset of images with people in real environments, annotated with their apparent emotions. Since the photos are collected from real-life settings, the images' variations are higher than other common datasets, such as CK+ and KDEF. This dataset includes facial occlusion (usually with a hand), partial faces, low-contrast images, and eyeglasses.

An extended list of 26 emotion categories is defined in this dataset to annotate the images, combined with three standard continuous dimensions: Valence, Arousal, and Dominance. Rather than recognizing emotion categories, several new studies on facial expression employ the aspects of the VAD Emotional State Model to depict emotions. The VAD model describes emotions using three numerical dimensions:

- Valence (V): A scale evaluates how pleasant or pleasant a feeling is, from negative to positive.
- Arousal (A): A scale that assesses a person's level of agitation, ranging from nonactive/calm to agitated/ready to act.
- Dominance (D): A scale that evaluates a person's amount of control over a situation, ranging from submissive/non-control to dominant/in-control [22].

The pictures in the EMOTIC dataset are mostly from well-known datasets such as MSCOCOC [24] and ADE20K [25]. The EMOTIC dataset consists of 18316 images with 23788 people annotated.

3.2.2. KDEF

The Karolinska Directed Emotional Faces (KDEF) [5] is one of the most widely used human facial expressions databases. KDEF is a collection of 4900 photographs depicting human face emotions. There are 70 people in the photo collection, each with a different emotional expression. Each emotion is examined from five distinct perspectives.

3.2.3. CK+ Dataset

Cohn-Kanade (CK) plus is the extended version of regular CK, which covers the shortcomings of its previous versions. CK+ has 593 sequences and 123 subjects, which is 22% more sequences and 27% more subjects than the original CK. Participants of CK+ dataset range between 18 and 50 years old. They were told to show 23 facial expressions consisting of single and multiple action units. The results of these sequences and subjects are distributed over seven different emotion categories that we are trying to detect.

3.3. Convolutional Neural Network

In this paper, we use Convolutional Neural Networks to detect emotions on our dataset. CNNs are the most popular architecture for image classifications. Pre-trained VGG-16 and Resnet-50 are customized to classify ten different categories of emotions. Customization of the VGG-16 and Resnet-50 are done by altering the classification part of the network. We added two fully-connected layers to produce ten outputs. Each output represents the probability of the image belonging to a specific category. No changes have been made to the feature-extraction architecture. In the next section, more detail is given about VGG and ResNet.

3.3.1. Residual Networks

After the first CNN-based architecture (AlexNet), which won the ImageNet 2012 competition, subsequent winning architectures use more layers in a deep neural network to minimize the error rate. The Residual Network [26] has a large number of layers. As the number of layers grows, the gradient vanishing problem arises. This issue changes the gradient value to either 0 or too large, which prevent the system to learn. Thus, as the number of layers increases, the training and test error rate also increases. ResNet resolves the vanishing/exploding gradient problem by adding the input features to the output. Fig. 2 demonstrates the residual block in the ResNet architecture.

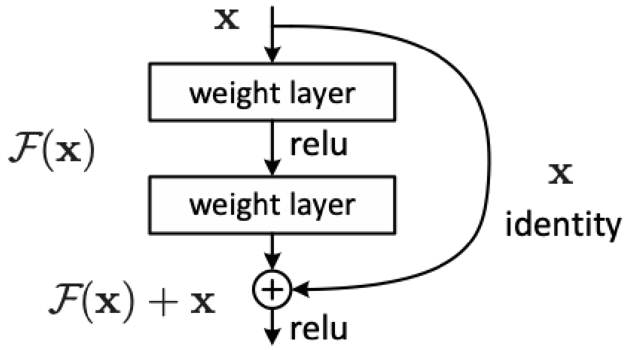


Figure 2. Residual Block in ResNet-50 architecture [27]

3.3.2. VGG

VGG-16 outperforms AlexNet by replacing large kernel filters sequentially (11 and 5 in the first and second convolutional layers, respectively) with numerous 3x3 kernel filters [27].

Rather than having a large number of parameters, VGG-16 employs 3x3 convolution filter layers with a stride 1. Also, the padding and maximum pooling layer with 2x2 filters and stride 2 remain the same. The convolution and max pool layers are placed in the same way throughout the design. At the end of its architecture, it has two fully-connected layers. The output is then followed by a softmax. The conv1 layer receives a 224 by 224 coloured image as input [28].

The features are extracted using convolutional layers with the smallest feasible dimensions: 3x3 to capture left/right, up/down, and centre of images. In one of the VGG variances, an extra 11 convolution filters are added, which may be regarded as a linear change to the input channels followed by non-linearity. The convolution spatial padding is set to 0 and the convolution stride is set to 1 pixel. After convolution, the spatial resolution of the layer input is preserved, i.e. the padding is 1-pixel for 33% of the convolutional layers. Spatial pooling is done via five max-pooling layers that follow part of the convolutional layers (not all the convolutional layers are followed by max-pooling). Max-pooling is done with stride 2 across a 2x2 pixel frame [29].

Following a stack of convolutional layers of varying depth in various designs, three Fully-Connected (FC) layers are added. 4096 channels are included in the first two FC layers. The last FC layer has 1000 channels since ImageNet dataset contains 1000 classes. The last layer performs as a softmax layer.

Rectified Linear Unit (ReLU) non-linearity is present in all hidden layers. Local Response Normalization (LRN), which does not enhance performance on the ILSVRC dataset but increases memory usage and computation time, is also included in none of the networks.

4. EXPERIMENT AND RESULTS

The dataset used in this work is a combination of three different datasets (CK+, EMOTIC, and KDEF), each of them having their unique features. Firstly, the augmented dataset is trained on a deep neural network, called FaceNet, to extract features from images of a person's face and detect the face. After the face images are detected, they should be augmented by image transformation to be fed to the input of emotion recognition networks. The final dataset consists of cropped, rotated, and horizontally-flipped images of the original dataset. After splitting the total images into training and testing 70% and 30%, respectively, the distribution of seven emotions are shown in Fig. 3 and Fig. 4.

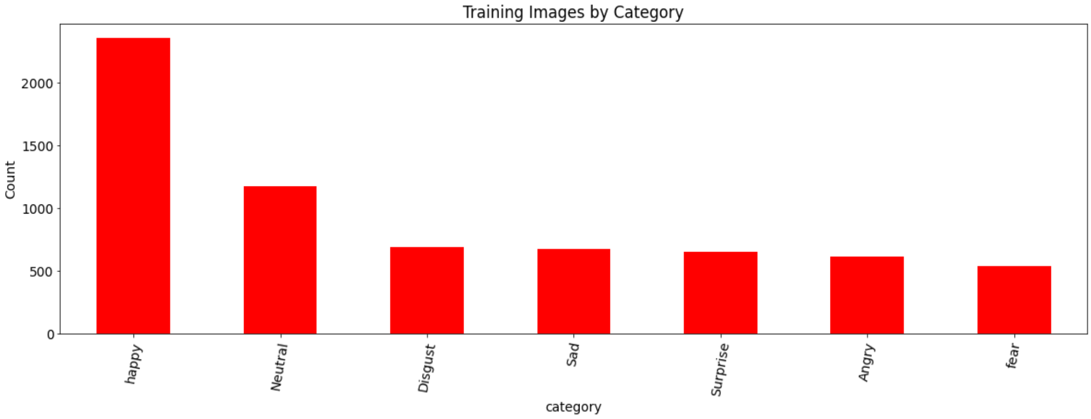


Figure 3. Distribution of training images in seven emotion categories.

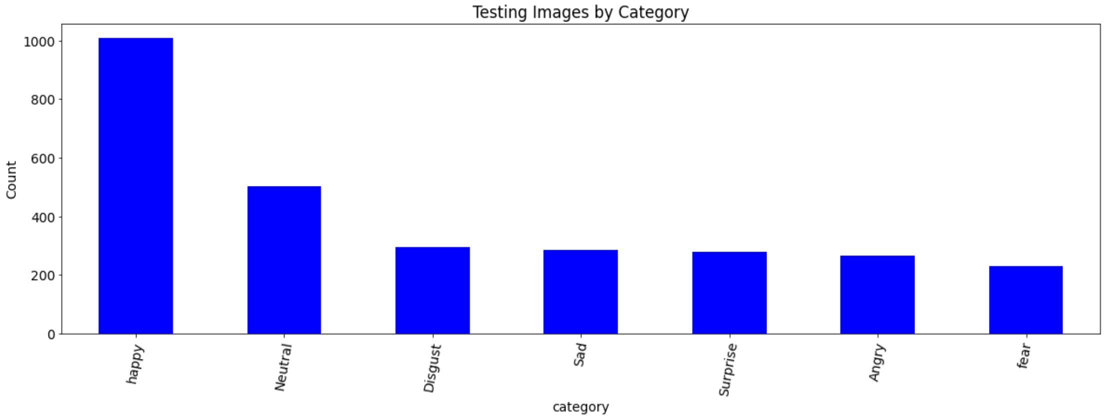


Figure 4. Distribution of testing images in seven emotion categories.

FaceNet, which is a Google-developed facial recognition system that obtains state-of-the-art results on various face identification benchmark datasets in 2015 [30], is used in this work to extract the faces. After extracting the face images, our pre-processed dataset is trained on two different convolutional neural network architectures: VGG-16 and ResNet-50.

As shown in Fig. 5, the accuracy of 52.49% is reached on the VGG-16 architecture. Moreover, the related confusion matrix is show in Fig. 6. Confusion matrix is a performance evaluation metric for machine learning classification problem. The actual target values and predicted values are compared using a confusion matrix. Labels are shown from 0 to 6, which maps to Anger, Disgust, Neutrality, Sadness, Surprise, Fear and Happiness. As illustrated, "Happiness" images are significantly higher in volume compared with other emotion categories. This explains the more accurate prediction of the architecture for this category shown in the confusion matrix as in Fig. 6.

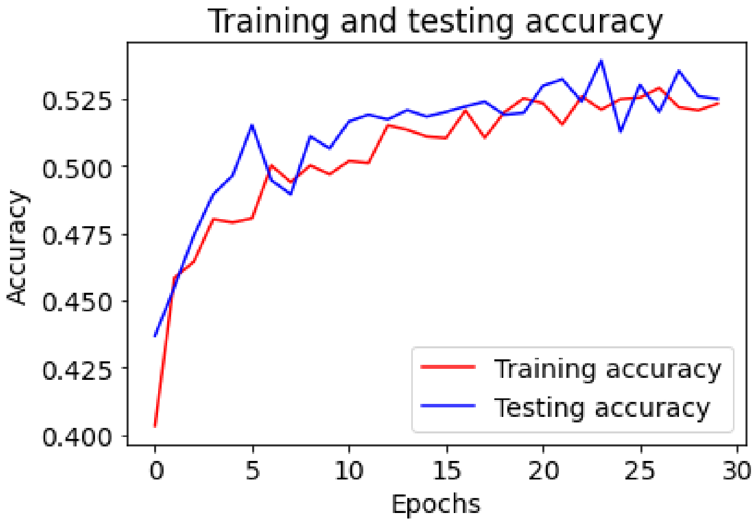
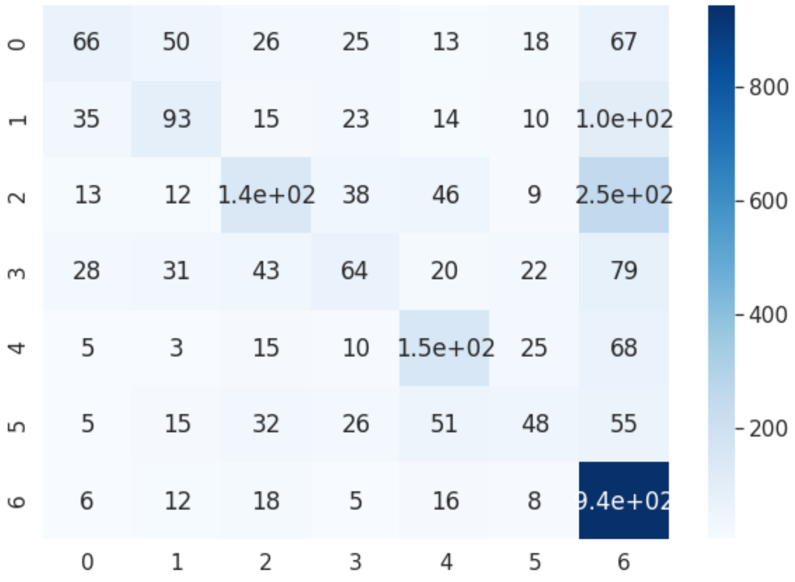


Figure 5. VGG-16- Training and testing accuracy with 30 epochs.



	0	1	2	3	4	5	6
0	66	50	26	25	13	18	67
1	35	93	15	23	14	10	1.0e+02
2	13	12	1.4e+02	38	46	9	2.5e+02
3	28	31	43	64	20	22	79
4	5	3	15	10	1.5e+02	25	68
5	5	15	32	26	51	48	55
6	6	12	18	5	16	8	9.4e+02

Figure 6. VGG-16- Confusion matrix.

As shown in Fig. 7, the accuracy of 55.63% is reached on the ResNet-50 architecture. The related confusion matrix is show in Fig. 8. As illustrated, "Happiness" images are significantly higher in volume compared with other emotion categories. This explains the more accurate prediction of the architecture for this category shown in Fig. 8.

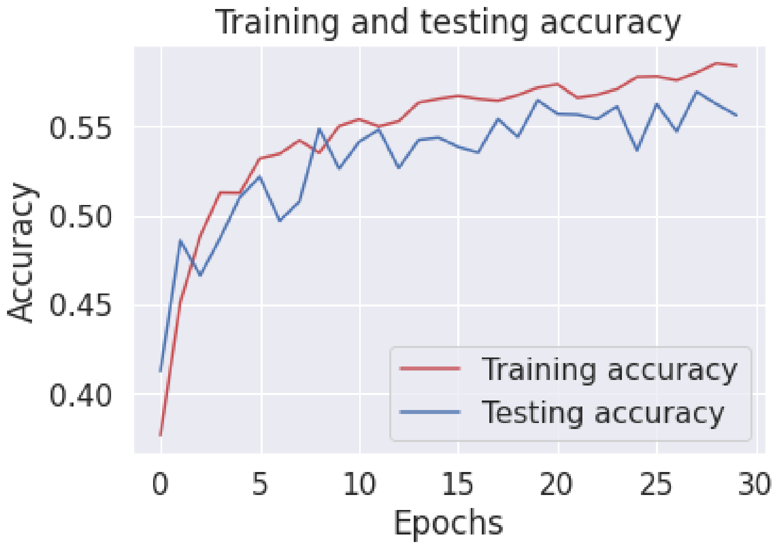


Figure 7. ResNet-50- Training and testing accuracy with 30 epochs.

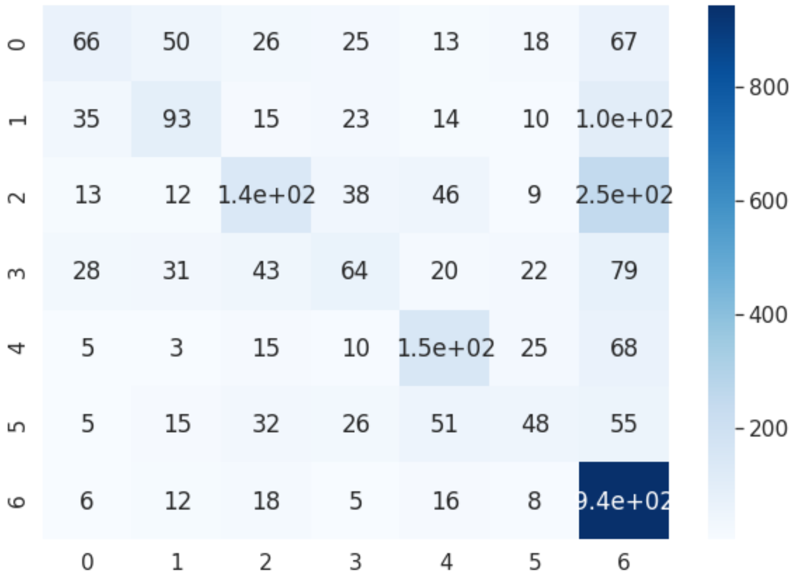


Figure 8. ResNet-50- Confusion matrix.

The above confusion matrices show high off diagonal values, which is due to class imbalance. To overcome this issue, a combination of alternative approaches called Synthetic Minority Oversampling Technique (SMOTE) and [31] is used.

In SMOTE, the minority class is over-sampled by creating “synthetic” examples instead of replacing the over-sampled examples [31]. The new samples are duplicated based on the Euclidean distance of each data and the minority class nearest neighbours. Therefore, the generated examples are different from the original minority class and provide additional information. This is useful for the system to learn the model.

In Tomek Link approach observations from the majority class are removed. This is also considered as an enhancement of Nearest-Neighbor Rule (NNR) [32]. This method uses NNR to select the pair of examples that fulfill specific properties. One of the advantages of this method is that it removes the data from the majority class that has the lowest Euclidean distance with the minority class data, therefore make it less ambiguous to detect the emotion.

For better comparison, we have shown how VGG-16 and ResNet-50 improved in Fig.9 and Fig. 10. The Y-axis shows the accuracy of the architecture over a certain emotion category. The X-axis represents the emotion categories from 0 to 6 mentioned previously. The enhanced dataset gives priorities to the emotion categories with less data.

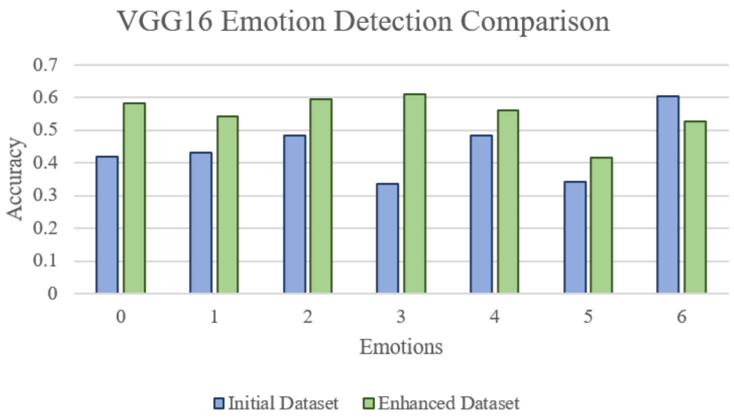


Figure 9. VGG-16 Emotion Detection Comparison.

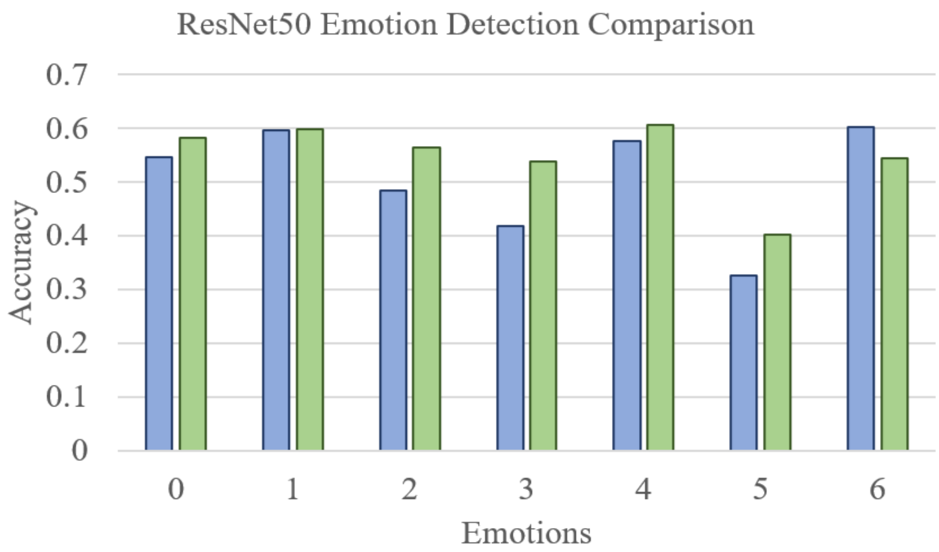
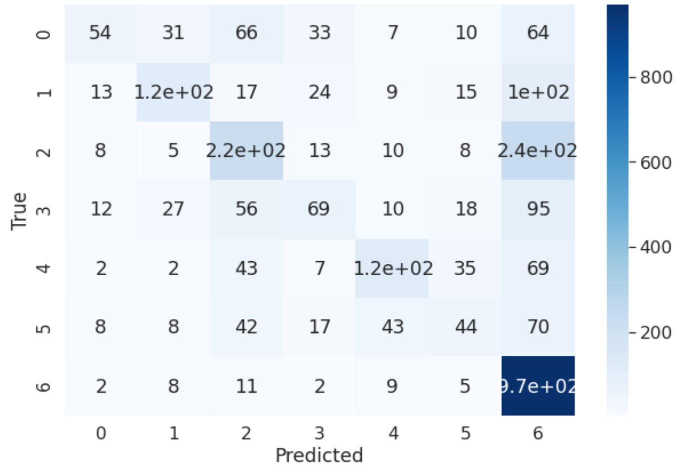


Figure 10. ResNet-50 Emotion Detection Comparison.



	0	1	2	3	4	5	6
0	54	31	66	33	7	10	64
1	13	1.2e+02	17	24	9	15	1e+02
2	8	5	2.2e+02	13	10	8	2.4e+02
3	12	27	56	69	10	18	95
4	2	2	43	7	1.2e+02	35	69
5	8	8	42	17	43	44	70
6	2	8	11	2	9	5	9.7e+02

Figure 11. VGG-16- Confusion matrix for enhanced dataset.

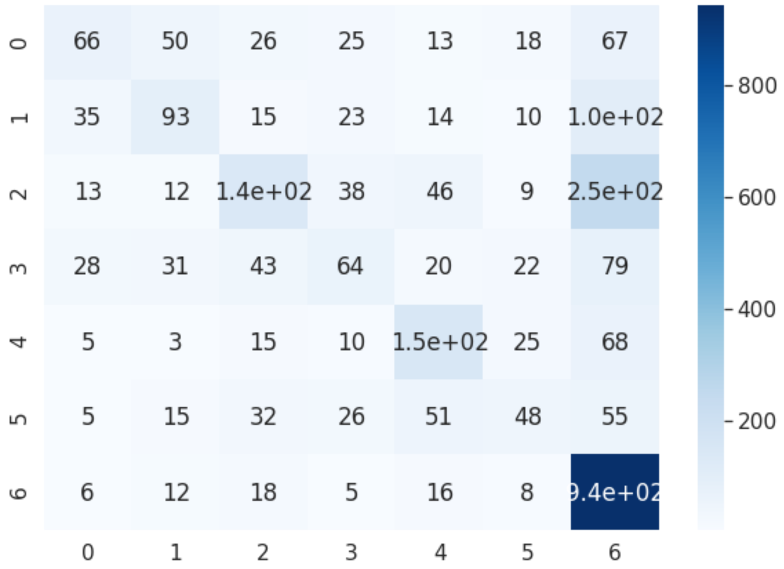


Figure 12. ResNet-50- Confusion matrix for enhanced dataset.

After modifying our dataset, using SMOTETomek approach, improvement over the initial method is observed. These improvements are reported over both architectures in Fig. 11 and Fig. 12. Comparing the enhanced confusion matrices with the initial confusion matrices, we can observe more consistency on seven emotion categories. Still, "Happiness" emotion distinction dominates other emotion detection.

Comparing the two confusion matrices (Fig.6 and Fig. 11), shows that our architecture produces more accurate results. This confirms the effectiveness of the proposed method compared with the initial one.

5. Conclusions

This work provides the implementation of facial emotion recognition based on two deep learning algorithms, VGG-16 and ResNet-50. From a technical point of view, this work has served to clearly demonstrate the advantage of using a balanced and enhanced dataset including almost the same number of examples in each class. The class imbalance data problem is also tackled using a combination of oversampling and undersampling techniques, called SMOTETomek. We have shown that VGG-16 and ResNet-50 can improve from about 50% up to 60.16% and 60.71% respectively. Future research can consider different architectures and fine-tuning hyper parameters.

References

[1] L. Tereikovska, I. Tereikovskyi, S. Mussiraliyeva, G. Akhmed, A. Beketova, and A. Sambetbayeva, "Recognition of emotions by facial Geometry using a capsule neural network," *Int. J. Civ. Eng. Technol.*, vol. 10, no. 03, pp. 1424–1434, 2019.

[2] P. Ekman, "FACIAL EXPRESSION Edited by An imprint of The Institute for the Study of Human Knowledge," 1973.

[3] A. Kelly, "Facial expression," *Talkabout*. pp. 61–70, 2019, doi: 10.4324/9780429427251-6.

[4] H. Facial, "Book Reviews," no. 1994, pp. 1187–1194, 1996.

[5] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-Visual Emotion Recognition in Video Clips," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 60–75, 2019, doi:

10.1109/TAFFC.2017.2713783.

[6] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, 2012, doi: 10.1109/T-AFFC.2011.37.

[7] Harappa, "Types Of Emotions." 2020.

[8] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial Expression Recognition Based on Facial Components Detection and HOG Features," 2014.

[9] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *IEEE International Conference on Image Processing 2005*, 2005, vol. 2, pp. II–370, doi: 10.1109/ICIP.2005.1530069.

[10] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 2, pp. 568–573 vol. 2, doi: 10.1109/CVPR.2005.297.

[11] J. Whitehill and C. Omlin, "Haar features for FACS AU recognition," *7th Int. Conf. Autom. Face Gesture Recognit.*, pp. 5 pp. – 101, 2006.

[12] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network," *Sensors*, vol. 21, no. 9, 2021, doi: 10.3390/s21093046.

[13] and W. V. F. Ekman, Paul, "Facial action coding system," *Environ. Psychol. Nonverbal Behav.*, 1978.

[14] H. C. William and W. T. M. Liao, "Facial emotion recognition with transition detection for students with high-functioning autism in adaptive e-learning," *Soft Comput.*, no. 1, 2017, doi: 10.1007/s00500-017-2549-z.

[15] S. Minaee, A. Abdolrashidi, and Y. Wang, "An Experimental Study of Deep Convolutional Features For Iris Recognition Electrical Engineering Department , New York University , Computer Science and Engineering Department , University of California at Riverside," *Signal Process. Med. Biol. Symp.*, 2016.

[16] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," 2016, doi: 10.1109/WACV.2016.7477450.

[17] D. Aneja, A. Colburn, G. Faigin, L. Shapiro, and B. Mones, "Modeling Stylized Character Expressions via Deep Learning," in *Computer Vision -- ACCV 2016*, 2017, pp. 136–153.

[18] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial Expression Recognition via a Boosted Deep Belief Network," *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1805–1812, 2014.

[19] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 92–105, 2011, doi: 10.1109/T-AFFC.2011.9.

[20] K. Schindler, L. Van Gool, and B. de Gelder, "Recognizing emotions expressed by body pose: A biologically inspired neural model," *Neural Networks*, vol. 21, no. 9, pp. 1238–1246, 2008, doi: 10.1016/j.neunet.2008.05.003.

[21] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 1960–1968, 2017, doi: 10.1109/CVPR.2017.212.

[22] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "EMOTIC: Emotions in Context Dataset," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2017-July, pp. 2309–2317, 2017, doi: 10.1109/CVPRW.2017.285.

[23] S. Ouyang, T. Hospedales, Y.-Z. Song, and X. Li, "A Survey on Heterogeneous Face Recognition: Sketch, Infra-red, 3D and Low-resolution." 2014.

[24] G. T. U. A. Colleges *et al.*, "Microsoft COCO," *Eccv*, no. June, pp. 740–755, 2014.

[25] B. Zhou *et al.*, "Semantic Understanding of Scenes Through the ADE20K Dataset," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 302–321, 2019, doi: 10.1007/s11263-018-1140-0.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." 2015.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[28] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." 2015.

[29] A. Savoiu and J. Wong, "Recognizing Facial Expressions Using Deep Learning."

[30] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and

clustering," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 815–823, 2015, doi: 10.1109/CVPR.2015.7298682.

[31] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[32] A. Elhassan and Al-Mohanna, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," 2017.