Improved Deep Convolutional Neural Network with Age Augmentation for Facial Emotion Recognition in Social Companion Robotics

*Steven Lawrence[1]

*Taif Anjum[2]

Amir Shabani[3]

{Steven.Lawrence,Taif.Anjum}@student.ufv.ca,
Amir.Shabani@ufv.ca

[1,2,3] Intelligent Systems and Computing Research Group,

School of Computing, Univ. of the Fraser Valley, BC, Canada

*: Equal Contributions

Abstract

Facial emotion recognition (FER) is a critical component for affective computing in social companion robotics. Current FER datasets are not sufficiently age-diversified as they are predominantly adults excluding seniors above fifty years of age which is the target group in long-term care facilities. Data collection from this age group is more challenging due to their privacy concerns and also restrictions under pandemic situations such as COVID-19. We address this issue by using age augmentation which could act as a regularizer and reduce the overfitting of the classifier as well. Our comprehensive experiments show that improving a typical Deep Convolutional Neural Network (CNN) architecture with facial age augmentation improves both the accuracy and standard deviation of the classifier when predicting emotions of diverse age groups including seniors. The proposed framework is a promising step towards improving a participant's experience and interactions with social companion robots with affective computing.

1. Introduction

Social companion robots [1] are receiving a lot of attention for the engagement of different age groups, including toddlers and seniors. More specifically, companion robots such as Miro-e [2] could play a significant role in compensating for the shortage of caregivers in the long-term care facilities which became more apparent during the COVID-19 pandemic [3]. Incorporating/leveraging technologies such as emotion recognition could both make the engagement more interactive and also provide more insights in the health and mental state of the residents.

Facial emotion recognition (FER) has been a long-standing research problem in computer vision and machine learning (ML) communities [4, 5]. Standard ML algorithms such as Support Vector Machines (SVM) and its Kernelized variations have been deployed significantly for different image/video classification problems [6, 7], including FER [4]. The new trend of using Deep Learning with CNNs and their variations has shown to be more effective, especially when large sample data is available and transfer learning [8] is of interest.

There are several publicly available FER datasets [9, 10, 11]. However, our extensive assessment shows that these datasets are not representative of our target group, i.e. seniors. As a result, the datasets are not diverse enough to be appropriate for our application of FER for social companion robotics. It is also worth noting the privacy in data collection from seniors is a very sensitive topic. This leads us to consider remedies such as data augmentation as an avenue to generate synthetic samples to include this age group in a dataset and have a better representative of the application domain. More specifically, data augmentation has been widely utilized to address the challenge of imbalanced datasets and also synthetic data generation when data collection is expensive or not readily accessible. For example, variations of Generative Adversarial Networks (GANs) [12] have been shown to successfully generate high-quality images over recent years [13].

Using face aging with the Identity-Preserved Conditional GAN (IPCGANs) [13] framework, we propose to extend the age range in the current FER datasets. This means we generate realistic samples for the fifty plus age group to be added to the training. Adding the new samples to the original dataset shows significant improvement in both selected datasets and also two state-of-the-art learning frameworks (i.e., MobileNet, Deep CNN) resulting in an improved deep learning framework for senior facial emotion recognition applications. Data augmentation works as a regularizer and reduces overfitting during training and hence improves accuracy. To the best of our knowledge, our framework is unique and is different from works such Xinyue et al. [14] where they address the issue of imbalance sample size among different classes in a dataset by augmenting the minority classes. Our focus is using data augmentation to increase the age diversity in a dataset which is a critical issue for our application of senior emotion recognition using social companion robots.

The rest of the paper is organized as follows. Section 2 describes in more detail the methodologies of two state-of-the-art classifiers (i.e, MobileNet and Deep CNN). Section 3 explains different FER datasets and in more detail two widely used ones (i.e., CK+ and RAVDESS). Section 4 provides the details of the experimental setup and the results. Section 5 discusses the main observations from five different experiments using the two classifiers and two datasets. Section 6 concludes the paper with the highlights of the findings and future directions.

2. Methodologies

Convolutional Neural Networks (CNNs) are the base of deep learning frameworks for FER in several studies with promising results [15, 16]. We used two CNN architectures for our experiments due to its promising success rate in different classification applications. Our Deep CNN classifier has six convolutional layers, three max-pooling layers, and two fully connected layers; all layers use Exponential Linear Unit (ELU) as their activation function. The output layer has nodes equal to the number of classes with a softmax activation function. Dropouts and batch normalizations were used at regular intervals to avoid overfitting. We also use MobileNet [17], a lightweight Deep CNN by Google, as our other architecture due to its efficiency. MobileNet is faster than many popular CNN architectures such as AlexNet, GoogleNet, VGG16, and SqueezeNet while having similar or higher accuracy [17]. The MobileNet classifier leverages transfer learning by using pre-trained weights from imageNet [18]. A fully connected output layer was added with nodes equal to the number of classes and softmax is used as the activation function. We used both Deep CNN and MobileNet classifiers with implementations from [19] in which the Nadam optimizer was used with a learning rate of 0.001 and compiled using categorical cross-entropy as the loss function for better classification. Two callbacks were used 'early stopping' to avoid overfitting and 'ReduceLRonPlateau' for reducing the learning rate when the learning stops improving.

3. Datasets

Many datasets are available for FER, namely Cohn-Kanade extended (CK+) [10], Japanese Female Facial Expression (JAFFE) [11], Facial Expression Recognition-2013 (FER-2013) [12], and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [20]. JAFFE is not racially diverse, FER-2013 lacks information such as age range, racial demographic, and other particulars because it has been collected from google images. The RAVDESS dataset consists of 24 professional actors with an age range from 21 to 33, out of which 12 are males and 12 are females with a mean age of 26 and a standard deviation of 3.75 years. Participants were mostly Caucasians and others were East Asian or mixed. CK+ has been widely used due to its diversity in sample data[21]. It includes the facial behavior of 210 adults, where the age of participants ranges from 18 years to 50 years, 69% female, 81% Euro-American, 13% Afro-American, and 6% other groups. To examine our hypothesis (i.e, the role of age augmentation), we needed a dataset that has a diverse age range and another one that is strictly age biased. This contrast in the datasets will help us to have more meaningful data to draw relevant conclusions. To this end, we chose two datasets, CK+ for its diversity in age group and RAVDESS for it being a very focused age group.

Data preprocessing was an integral part of our experiments. Each image was cropped to their face using face detection with OpenCV's Haar Cascade Classifiers [22]. The images were resized to 48x48 and converted to grayscale. For both datasets, the last three frames, referred to as the peak emotion, were used for each individual. For unbiased experimentations and fair comparisons, only the common classes between the two datasets were used, namely the six classes of Anger, Fear, Disgust, Sadness, Surprise, and Happy. After preprocessing, CK+ has 927 images and RAVDESS has 6,897. After applying face-aging augmentation (i.e., doubling the sample size), CK+ has 1854 images, and RAVDESS has 13,794 images. Figures 1 and 2 provide sample images of the RAVDESS and CK+ dataset after preprocessing the first row is the original samples and the second row is the augmented versions. Benchmark accuracies on CK+ for FER are 99.7% using Frame Attention Networks (FAN) [23], 98.6% using FaceNet2ExpNet [24] and 96.9% using MicroExpNet [25] a lightweight and fast architecture. We were unable to find benchmark accuracy on RAVDESS for FER as most studies have used it for speech emotion recognition. It should be noted that because of different pre-processings, implementations, frameworks, and applications, these results should be compared tentatively. For example, [23] uses all the frames in the video while we are only using the last three peak emotion frames which is less time complex and more applicable to our application of robotics with embedded/edge computing. Our main focus in using these datasets is to evaluate the role of face aging in the performance of the classifiers.

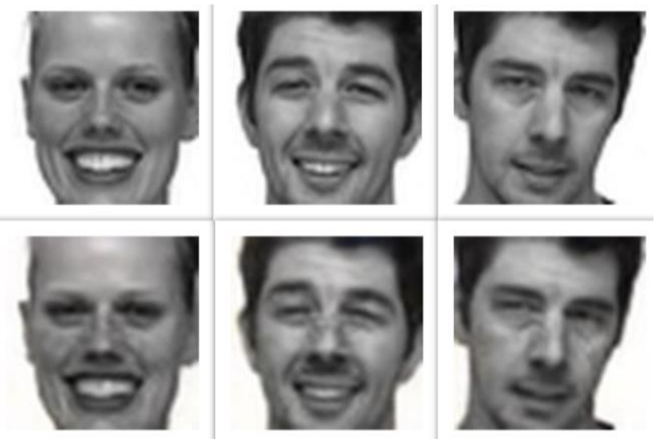

Fig. 1: RAVDESS dataset - Original faces (first row) vs. Age-augmented faces (second row).

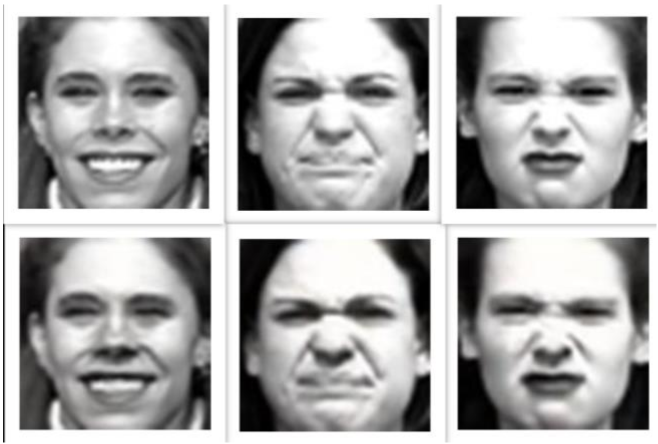

Fig. 2: CK+ dataset - Original faces (first row) vs. Age-augmented faces (second row).

4. Experiments & Results

We used the pre-trained face-aging model developed by Wang et al [13] which is referred to as Identity-Preserved Conditional GAN (IPCGANs) to generate our face-aging dataset. The model generated four images for age groups 20-30, 30-40, 40-50, and 50+. We only used images for the 50+ age group in our experimentations. After augmentation, both of the datasets doubled in size. For all our experiments, we split the datasets into three subsets using 30-fold stratified cross-validation, 60% for training, 20% for validation, and 20% for testing.

For a more concise representation, we are using the following notations for referring to different data samples and experiments:
Orig.: Contains only the original data from the dataset.
Aug.: Contains only the augmented data variants from dataset (excludes the original data).
Orig.+Aug.: Contains both the original data and the corresponding augmented variants.

Experiment 0: This experiment is our base case where we test the accuracy of our classifiers on the respective datasets with no augmentation done.
Training: 60% Orig.|**Validation**: 20% Orig.| **Testing**: 20% Orig.

Experiment 1: This experiment is to demonstrate the effect face aging has on the classifier's accuracy. After the dataset has been

split into the appropriate subsets, the testing data is exchanged with only corresponding augmented variants.
Training: 60% Orig.| **Validation**: 20% Orig.|**Testing**: 20% Aug.

Experiment 2: This experiment is to simulate a testing pool that contains all age groups. The testing group contains the 20% original subset of the dataset along with the matching augmented variants.
Training: 60%Orig.|**Validation**: 20%Orig.|**Testing**: 20% Orig.+Aug.

Experiment 3: This experiment is to demonstrate the effects age augmentation has on the accuracy of the classifier when tested on its original data.
Training: 60% Orig.+Aug.|**Validation**: 20% Orig.|**Testing**: 20% Orig.

Experiment 4: This experiment is to demonstrate the results of age augmentation when used as a component in all subsets of the classifiers.
Training: 60% Orig.+Aug. | **Validation**: 20% Orig.+Aug.| **Testing**: Orig.+Aug.

Table 1 contains the average test accuracy with standard deviation over 30 runs for each of our experiments where experiment 0 is the results of the baseline and experiment 4 is the result of our proposed solution with age data augmentation.

Table 1: Test accuracy (Avg.± Std.) over 30 runs on different datasets using different classifiers.

Dataset	**Experiment**	**MobileNet**	**Deep CNN**
CK+	**0 (Baseline)**	**94.89 ± 5.57 %**	**89.87 ± 3.50 %**
	1	92.53 ± 3.94 %	87.58 ± 4.40 %
	2	94.10 ± 2.76 %	88.92 ± 4.29 %
	3	97.74 ± 1.70 %	96.13 ± 2.88 %
	4 (Ours)	**99.10 ± 0.84 %**	**98.28 ± 1.06 %**
RAVDESS	**0 (Baseline)**	**89.07 ± 2.67 %**	**91.99 ± 2.14 %**
	1	83.58 ± 3.70 %	86.02 ± 2.66 %
	2	86.96 ± 2.80 %	88.67 ± 2.43 %
	3	94.11 ± 1.81 %	95.96 ± 1.24 %
	4 (Ours)	**95.12 ± 1.35 %**	**96.97 ± 1.27 %**

Figures 3 and 4 are visual representations of how the test accuracy increases for the classifiers as we advance from experiment 0 to 4. This demonstrates how the classifiers react as we introduce more augmented data.

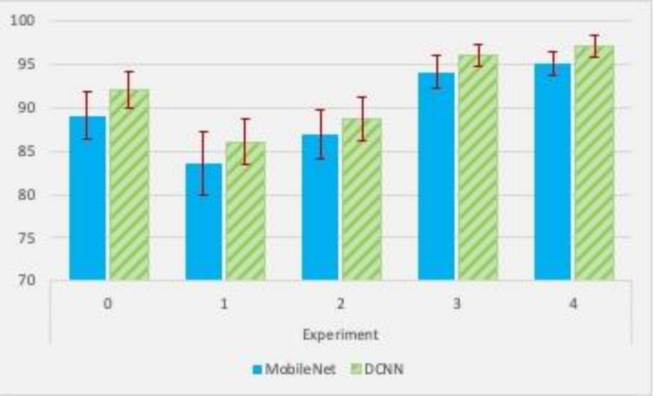

Fig. 3: Testing accuracy (Avg.± Std.) over 30 runs for different experiments on RAVDESS dataset.

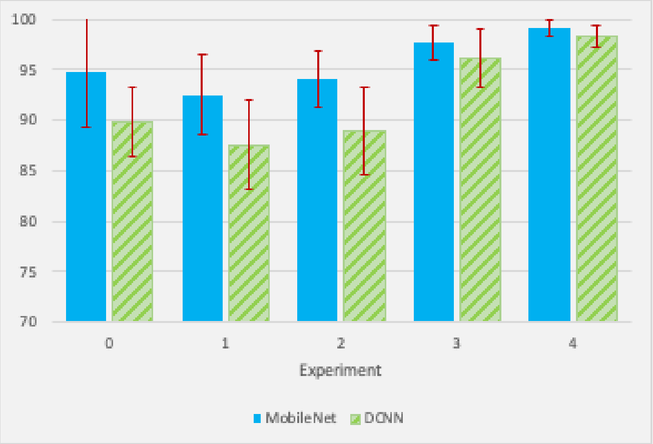

Fig. 4: Testing accuracy (Avg.± Std.) over 30 runs for different experiments on CK+ dataset.

Figure 5 shows that as more portions of the data are augmented from experiment 0 to experiment 4, the standard deviation for test accuracy significantly decreases. This produces a more reliable and precise classifier. By experiment 4, when the classifiers see only Orig.+Aug. data the standard deviation becomes almost the same for both classifiers.

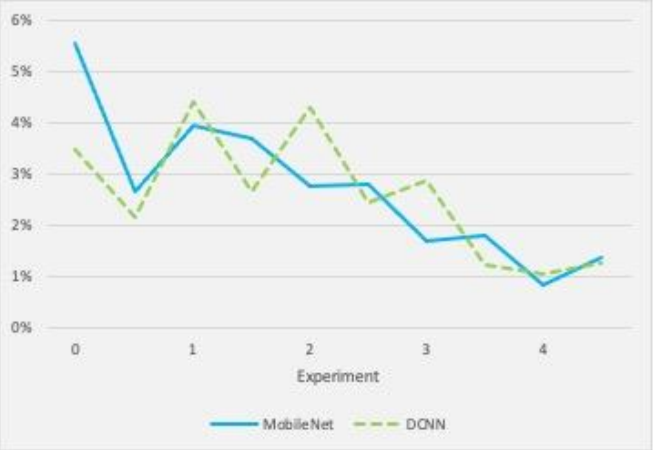

Fig 5. Standard Deviation (Avg) over 30 runs for MobileNet and Deep CNN By Experiment.

5. Discussion

In this section, we discuss our observations from different experiments. To provide more clarity it is worth mentioning when discussing experiments 1 through 4, by default, we will be drawing comparisons between experiment 0 which is the baseline. When referring to an increase or decrease this is with regards to experiment 0 (baseline) unless otherwise stated.

Experiment 1 has been designed to observe the role of age diversity and validate whether including aged data can affect the classifier's accuracy. In this experiment, the classifier trained using only original non augmented data, it was then tested strictly on age augmented data. The results of this experiment show that the accuracy of MobileNet decreases about 2.36% on CK+ while this decrease is 1.79% for Deep CNN. With RAVDESS, the MobileNet (Deep CNN) classifier displayed a 5.49% (5.97%) decrease. These results indicate the classifiers struggle to predict synthetically aged facial images which make sense because they haven't seen such samples during training. The accuracy reduction is less severe on CK+ as it has more diversity in the age range in the samples when compared to RAVDESS. The initial argument that aging can affect the classifier's accuracy is validated by the results of experiment 1.

Experiment 2 tests the classifiers on a subset of data consisting of original and age-augmented data to account for all age variations. The results show MobileNet's accuracy decreases by about 0.79% on CK+ while this decrease is slightly more for Deep CNN (0.95%). This decrease in accuracy is more on RAVDESS dataset, 2.11% for MobileNet, and 3.32% for Deep CNN. This experiment further proves the need to have age-diversified samples in the training phase to more accurately predict face-aging in FER.

Looking at the results of Experiment 1 and Experiment 2, we can observe that the age diversity is important in improving the accuracy and standard deviation, independent from the classifier model, i.e., the observation is consistent in both classifiers.

Experiment 3 evaluates the effect of including augmented data in the training stage as well. The two classifiers are trained on both original and augmented data. The trained classifier is then tested on only unseen original data to precisely see the effect of augmentation on the training process. As it can be seen in Table1, both MobileNet and Deep CNN are showing an increase in accuracy in both datasets when compared to the baseline. More specifically on CK+, MobileNet shows 2.85% improvement while this number is 6.26% for Deep CNN. And on RAVDESS, these numbers are 5.04% and a 3.97%, respectively. This experiment validates our hypothesis that having augmented data for training can significantly improve the classifier's accuracy and reduce the overall standard deviation which leads to a more reliable classifier.

Experiment 4 tests the effect age augmented data has on the classifier in its entirety. In a more realistic scenario, we wished to have real sample data of the aged group (+50) for this testing. In absence of this data, we generated augmented samples for testing data and included it in the set. The two classifiers are trained, validated, and tested on the full augmented dataset containing all original data and their augmented variants. For CK+, we receive a 4.21% and 8.41% increase in the average test accuracy using MobileNet and Deep CNN, respectively. For RAVDESS, we receive a 6.05% and a 4.98% increase. This experiment shows that including age augmented data during training can significantly increase a classifier's accuracy on both original and augmented data. This validates our hypothesis that using face aging augmentation can help the classifiers better predict emotions of different age groups. This improvement is more on RAVDESS as is less age-diversified. However, the improvement on already age-diversified CK+ is also statistically significant. It is also worth noting that Deep CNN is benefiting more from the regularization by adding age-augmented data and hence reduces overfitting during learning.

When we compare figures 3 and 4 we see that figure 4 for CK+ has a higher overall test accuracy. This is the result of CK+ already being a semi-age diverse dataset, whereas RAVDESS is not.

Our experiments show that MobileNet performs better on CK+ whereas Deep CNN performs better on RAVDESS. MobileNet incorporates transfer learning using imageNet, that allows the classifier to be more diverse and generalized. In contrast, Deep CNN does not have the transfer learning component, it only uses the given dataset during training, as hence it is less generalized. This is why MobileNet performs better on a more diverse dataset such as CK+ compared to Deep CNN performing slightly better on a less diverse dataset of RAVDESS.

6. Conclusions & Future Work

In this paper, we proposed a solution in addressing the lack of age diversity in facial emotion recognition datasets which becomes more of a severe issue when being used in social companion robots towards affective computing [26]. Datasets are typically biased in some manner, whether it be age, ethnicity, gender, or any other characteristic. Collecting data can be expensive, time-consuming, and difficult, therefore all ventures cannot be explored and collected from. To bridge this gap, data augmentation has been a successful tool. In our application, age demographic is the key and the (training) dataset should have a diverse age group to be relevant and effective when tested for the senior age group (50+). We incorporated a data augmentation scheme using IPCGAN [13] and improved the accuracy of two different Deep Learning frameworks (MobileNet and Deep CNN) on two different datasets. Our comprehensive experiments show that the accuracy of existing solutions, when tested on aged faces decreases significantly. When our face-aging augmentation has been added to the training set, the classifier's accuracy in predicting emotions of different age groups has been improved and the standard deviation of the classifier reduced. This is because data augmentation works as a regularizer and reduces overfitting during training and hence improves testing accuracy. This validates the proposition that the age gap in biased datasets could be bridged through proper data augmentation and hence solutions be more generalized and applicable.

Our future work will include cross-dataset experimentation to provide an even more generalized solution using data augmentation. This also includes evaluating the effects of transfer learning from one classifier to another. In the near future, we extend our solution towards affective computing with deployment on an embedded computing device to enable our social companion robot (Miro-e) to be able to respond in real-time to the participant's emotions and improve their experience and interaction. A longer-term vision is collaboration with our health science research team members and assess the effectiveness of our solution in a clinical setting in long-term care facilities.

References

[1] C. A. Cifuentes, M. J. Pinto, N. Céspedes, and M. Múnera, “Social Robots in Therapy and Care,” *Current Robotics Reports*, 1-16, 2020.

[2] T. J. Prescott, B. Mitchinson, and S. Conran, "Miro: An animal-like companion robot with a biomimetic brain-based control system," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 50-51. 2017.

[3] C. H. Chu, S. Donato-Woodger, and C. J. Dainton, “Competing crises: COVID-19 countermeasures and social isolation among older adults in long-term care,” in *Journal of Advanced Nursing*, 2020.

[4] B. C. Ko, “A brief review of facial emotion recognition based on visual information,” *sensors*, vol. *18* no. 2 , pp. 401, 2018.

[5] S. Li, and W. Deng, “Deep facial expression recognition: A survey,” in *IEEE Transactions on Affective Computing*, 2020.

[6] A. H. Shabani, J.S. Zelek, D.A. Clausi, “Multiple scale-specific representations for improved action classification”, in *Journal of Pattern Recognition Letters*, Jan. 2013.

[7] A. H. Shabani, J.S. Zelek, D.A. Clausi, "Improved Spatio-temporal Salient Feature Detection for Action Recognition," in The *British Machine Vision Conference*, pp. 1-12. 2011.

[8] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” in *Journal of Big data*, vol. 3, no. 9, 2016.

[9] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in *IEEE computer society conference on computer vision and pattern recognition-workshops*, IEEE, San Francisco, CA, USA, pp. 94-101, June 2010.

[10] M. J. Lyons, M. Kamachi, and J. Gyoba, “Japanese female facial expressions (JAFFE),” in the *database of digital images*, vol. *3*, 1997.

[11] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner and Y. Zhou, “Challenges in representation learning: A report on three machine learning contests,” *in the International conference on neural information processing,* Springer, Berlin, Heidelberg, pp.117-124, November 2013.

[12] I. J. Goodfellow and P. A. Jean and M. Mehdi, X. Bing and W. F. David, O. Sherjil, C. Aaron, and Yoshua Bengio, “Generative Adversarial Networks” in *Advances in neural information processing systems*, pp: 2672-2, 2014.

[13] Z. Wang, X. Tang, W. Luo, and S. Gao, “Face aging with identity-preserved conditional generative adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Shenzhen, China, pp. 7939-7947, 2018.

[14] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin, “Emotion classification with data augmentation using generative adversarial networks,” in *Pacific-Asia conference on knowledge discovery and data mining,* Springer, Cham, pp. 349-360, June 2018.

[15] M. Wafa, H. Wahida, “Facial emotion recognition using deep learning: review and insights,” in *Procedia Computer Science*, vol. 175, pp. 689-694, 2020.

[16] D. K. Jain, P. Shamsolmoali, and P. Sehdev, “Extended deep neural network for facial emotion recognition,” in *Pattern Recognition Letters*, vol. *120*, pp. 69-74, 2019.

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” in *arXiv, preprint arXiv:1704.04861*, 2017.

[18] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database”, in *IEEE conference on computer vision and pattern recognition*, Miami, FL, USA, IEEE. (pp. 248-255) June 2009.

[19] Gaurav Sharma, Facial Emotion Recognition, https://github.com/greatsharma/Facial_Emotion_Recognition, 2020.

[20] S. R. Livingstone, S. R, and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” in *PloS one*, vol. *13*, no. 5, e0196391, 2018.

[21] S. Li, and W. Deng, “Deep facial expression recognition: A survey,” in *IEEE Transactions on Affective Computing*, 2020.

[22] L. Cuimei, Q. Zhiliang, J. Nan and W. Jianhua, "Human face detection algorithm via Haar cascade classifier combined with three additional classifiers," *2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, Yangzhou, pp. 483-487, 2017.

[23] D. Meng, X. Peng, K. Wang, and Y. Qiao, “Frame attention networks for facial expression recognition in videos,” in *2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, pp. 3866-3870, IEEE, September 2019.

[24] H. Ding, S. K. Zhou, and R. Chellappa, “Facenet2expnet: Regularizing a deep face recognition net for expression recognition,” in *the 12th IEEE international conference on automatic face & gesture recognition, IEEE*, pp. 118-126. May, 2017.

[25] I. Cugu, E. Sener, and E. Akbas, “MicroExpNet: An Extremely Small and Fast Model For Expression Recognition From Face Images.” in *Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Istanbul, Turkey, pp. 1-6, 2019.

[26] Picard, Rosalind W. *Affective computing*. MIT press, 2000.