Investigation on Light-Weight Deep Learning Model for Emotion Recognition Using Facial Expressions

Su Yen Ding, Tong Boon Tang, *Senior Member, IEEE,* and Cheng-Kai Lu, *Senior Member, IEEE*

***Abstract*— Research findings have unveiled that facial expressions possess the ability to convey a variety of intense emotions. Hence, in this study, a deep-learning based approach, 2-Dimensional Convolutional Neural Network (2D CNN) for facial emotion recognition is proposed. The proposed network is running at least 47.28 times lesser number of parameters at 542,136, compared to the state-of-the-art (SOTA) network from RAVDESS dataset. The saving from reduced parameters is expected to translate into faster execution in real time. The proposed network scored accuracy of 92% and 94% that outperformed majority of the SOTA networks trained on RAVDESS and SAVEE dataset respectively, except one LSTM network from RAVDESS dataset that scored 98.90% in accuracy but with 116.5x higher number of parameters.**

***Clinical Relevance*— Highly accurate predictions from proposed lightweight architecture might aid the accessibility of lower computational power device to emotion recognition.**

I. Introduction

Facial expressions are often more intuitive in human daily interactions as they are a more direct reflection of emotion, if compared to body language. In general, emotion recognition by facial expressions involves either 2D face image or 3D video sequence with convolutional neural network (CNN) of different dimensions. In recent years, it is increasingly common to see high performance reported in various studies. Li *et al.* [1] reports an accuracy of 97.38% and 97.18% using CohnKanade (CK+) and Japanese female Facial Expression (JAFFE) dataset, respectively, trained on a simple 2D CNN consists of only two convolution layers. Work done by Li and Lima [2] recognizes emotion from facial expression at an accuracy of 95.39% using deep-CNN (DCNN) model ResNet50. Another research by Akhand *et al.* [3] delineated emotion at an accuracy of 96.51% and 99.52% with Karolinska Directed Emotional Faces (KDEF) and JAFFE dataset, respectively, using DCNN model DenseNet161.

Minaee *et al.* [4] proposed attentional convolutional neural network to perform emotion recognition and achieved 99.3% of accuracy using Facial Expression Research Group Database (FERG) dataset. Studies with high performance 2D CNN are well established, however, the efficiency of such a model is seldom being investigated. The main differences among the literature are that (a) some implemented face cropping and rotation strategy [1], (b) some performed background removal and face vector extraction [5], and (c) some adopted transfer learning with DCNN [3], [4]. The 3D CNNs also have similar high performance. For instance, Jeong [6] reported an accuracy of 99.21% using CK+ dataset and Lee [7] reported accuracy of 97.9% using BigFaceX dataset. However, CNN's input is dimensional-dependent, and it linearly increases with the computational cost for forward propagations, as well as the complexity. To keep the computational cost at its minimal, lightweight CNN is generally preferable, regardless of its input dimension, specifically for less complex problems [8]. A lightweight CNN typically comprised a lesser convolution layer count and ability to produce shallow feature map through each successive convolutions, resulting in less complex model [9]. Reduction in model complexity also refers to fewer number of parameters, hence lower computations required [10]. Faster execution is concurrently achieved through the minimization of model complexity [11]. By leveraging fewer number of parameters, lightweight CNNs can swiftly perform inference for predictions.

In facial emotional recognition, the datasets that comes with video sequences demands extra computational cost if every single frame is considered, which might also be inclusive of redundant frames (e.g., frames that does not have any facial representation). Thus, in addition to computational cost, extraction of face images is also needed to produce an input for the model, specifically for 2D CNN model. That can be realized through facial landmark detection algorithms [12]. Essentially, there are three commonly used facial landmark detection algorithms in Python, which are: 1) OpenFace that works with real time data video and performs facial landmark recognition simultaneously [13]. Other than that, this tool can estimate the head posture, recognize the facial action unit and capable of performing eye-gaze estimation for the face; 2) dlib that can be used for both facial detection and facial landmark detection [14]; 3) OpenCV that is an open-source library for image and video analysis that can be used for task related to facial detection and facial analysis [15]. According to [16], [17], the performance of OpenCV is better than dlib (shorter recognition time). When a face is detected, the coordinates of the facial landmark can be obtained and the positions of 68 points in human face can be predicted.

In this work, facial emotion recognition using deep learning-based approach using a tailored made 2D CNN layers is proposed. The prime objective is to produce a lightweight convolutional neural network while retaining the highest performance it possibly could. The proposed architecture is trained and validated with Multi-view Emotional Audio-Visual Dataset (MEAD), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Surrey Audio-Visual Expressed Emotion (SAVEE) dataset. The remaining part of this paper is structured as follows: Section II describes the methodology; section III discusses the result and is concluded in section IV.

*This study is funded in part by PETRONITA and the Ministry of Higher Education Malaysia HICoE grant to CISIR, UTP.

Su Yen Ding is with the Centre for Intelligent Signal and Imaging Research, University Teknologi PETRONAS, Perak, Malaysia (e-mail: su_22000257@utp.edu.my).

Tong Boon Tang is with the Centre for Intelligent Signal and Imaging Research, University Teknologi PETRONAS, Perak, Malaysia (e-mail: tongboon.tang@utp.edu.my).

Cheng-Kai Lu is with the Electrical Engineering Department, National Taiwan Normal University, Taipei City, Taiwan (e-mail: cklu@ntnu.edu.tw).

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

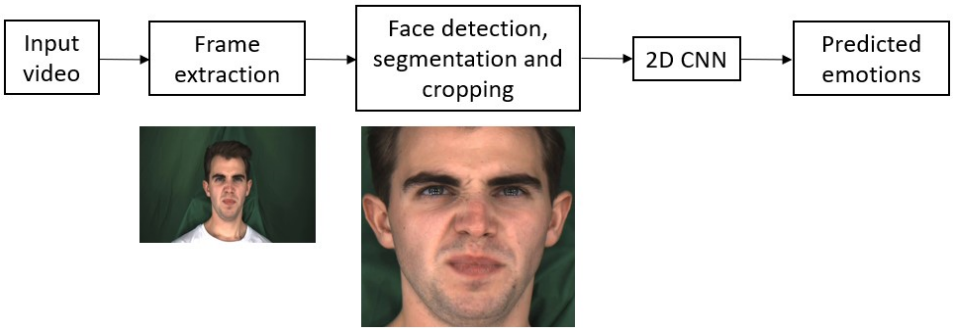

Fig. 1. Proposed methodology for emotion prediction using facial images.

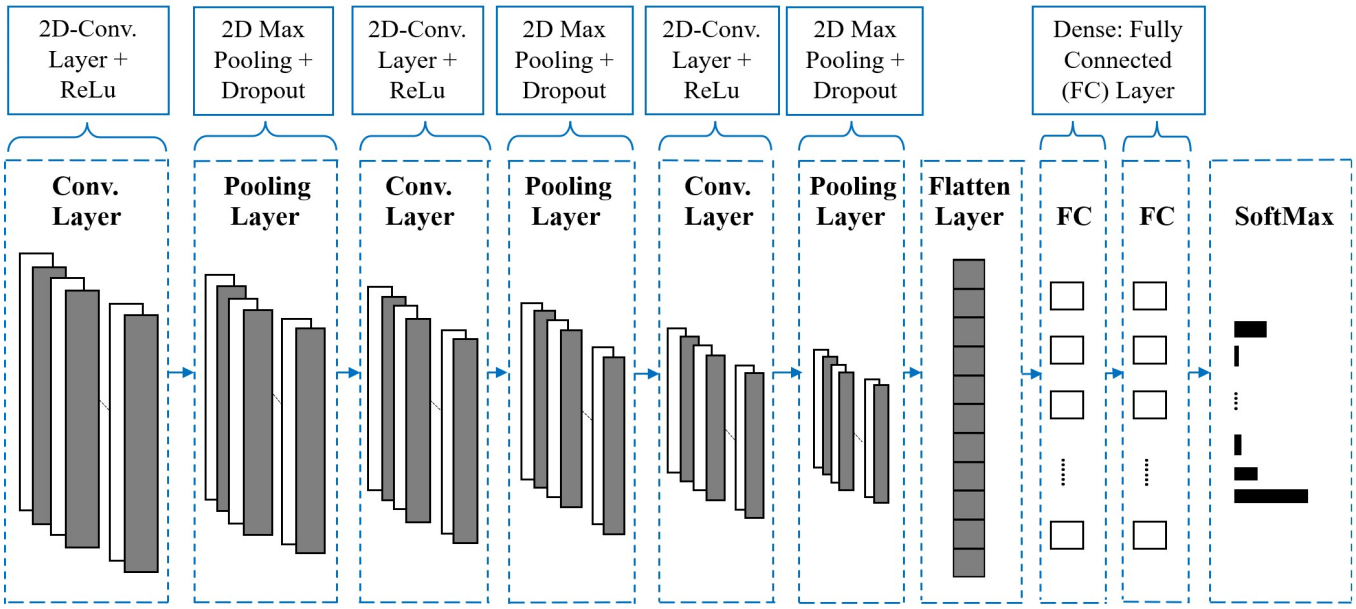

Fig. 2. Proposed 2D CNN architecture consisting of three convolution layers with each followed by successions of max pooling layer.

II. Methodology

The proposed methodology is described in Fig. 1. Firstly, the stream of video was pre-processed with frame extraction to dissect the stream of video into an image sequence (frames). Then, the extracted frames were pre-processed with a face detection algorithm and were further cropped to retain only the subject's face. The cropped images were then used as an input to the 2D CNN for the prediction.

A. Dataset

A total of three publicly available datasets were utilized in this study. The first dataset is the MEAD (total 168576 images) from [18] where the dataset consists of eight different sentiments (angry, disgust, contempt, fear, happy, sad, surprised, and neutral) at three different intensity levels each, except for neutral, with seven different viewing angles in a strictly controlled environment. The second dataset is the RAVDESS (total 17250 images) [19]. The dataset provides eight emotions (neutral, calm, happy, sad, angry, fearful, disgust and surprised) at two intensity level of 24 professional actors (12 females and 12 males). The assigned task is vocalizing two lexically matched statement in a neutral North American accent. The third and last dataset is the SAVEE (total 2880 images) [20] that comes with seven emotions (anger, disgust, fear, happiness, sadness, surprise and neutral) but only at one intensity level. Participants were required to perform recording based on 15 TIMIT sentences for each emotion as stimulant.

B. Pre-processing

The frame sequence was extracted from the video stream by taking the frame rate or frame per second (fps) into account. However, extracting all images for analysis would require higher computation resources. Therefore, only six frames were extracted from each video sequence [21]. After extracting the images, face detection and segmentation were applied. The dlib was used for facial detection and image cropping. The images undergone RGB-to-grayscale conversion before being scaled to the required input size of designed architecture, which is at 180 × 180.

C. 2D CNN for classification

After pre-processing, the images were fed into 2D CNN to perform classification. The classification of the data was based on the naming of the image file, which is based on the emotion type. Fig. 2 shows the proposed architecture. The facial images went through multiple Rectified Linear Unit (ReLU) activated convolutions, with pooling and dropout layers sandwiched in between. The convolutions have same size of filters at 3 × 3, which aims to capture more information about the low-level features of the image (e.g., edges, angles) at first convolution. Then, every convolution was followed by max pooling that shrunk the feature map every succession. By retaining the filter size at second and third convolutions, it is assumed that the model focuses only some parts of the face. Though, the number of filters is quadruple from the initial convolution to capture as many combinations of pattern as possible, for that focused part.

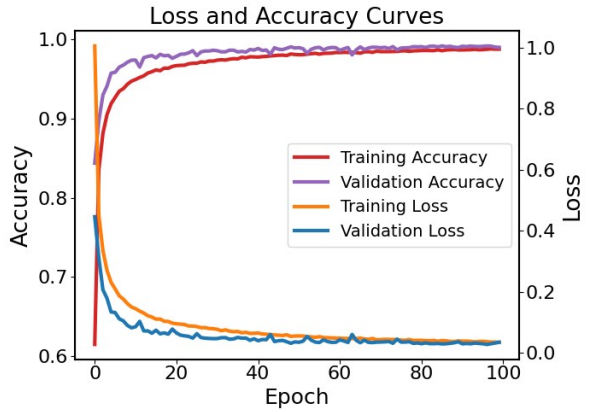

Fig. 3. Model's training and validation plots using MEAD dataset

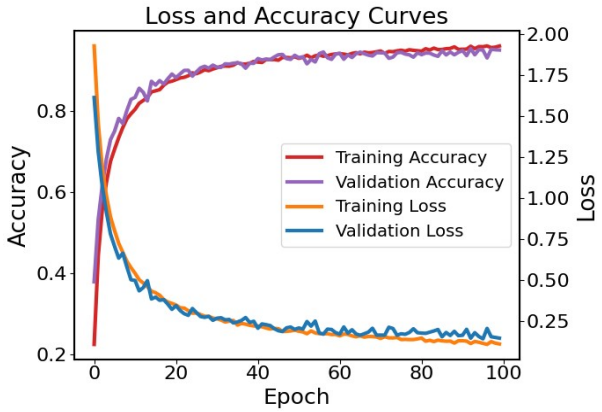

Fig. 4. Model's training and validation plots using RAVDESS dataset

The complete feature map was then flattened so that it could weigh into fully connected layers and classified by the SoftMax activation function.

D. Model training and performance evaluation

Each of the datasets was split into a training and testing set based on 80/20 rule where 80% of the data is used for training and 20% of the data is used for testing. Performance evaluation of the model was based on the accuracy and loss graph. Additionally, F1-score is calculated to provide a summary of the model's generalizability. It is generally expressed as (1).

$$F1 - score = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (1)$$

III. RESULTS

The hardware utilized for the experiments were 12th Gen Intel® Core™ i5-12500H CPU (~2.50 GHz), NVIDIA GeForce RTX3060 Laptop GPU (3840 CUDA cores) and 32GB of DDR5 RAM operating at 4800MHz. Concurrently, the implementation of models was performed using Python version 3.9.12 with Anaconda distribution, specifically conda version 4.12.0. Fig. 3 and 4 show the accuracy and loss graphs of the model's training and testing for MEAD and RAVDESS, respectively. Meanwhile, Fig. 5 and 6 are the graphs for SAVEE dataset before and after tuning the dropout layer and L1 regularizer. Based on the depicted graphs, the proposed model does not overfit the MEAD dataset, and RAVDESS dataset. While for SAVEE, it is obvious that the model suffers from overfitting, based on the diverging trend in Fig. 5. Overfitting is a common issue in model development where the model starts to memorize one group of features and becomes exceptionally well in predicting that group. This is an

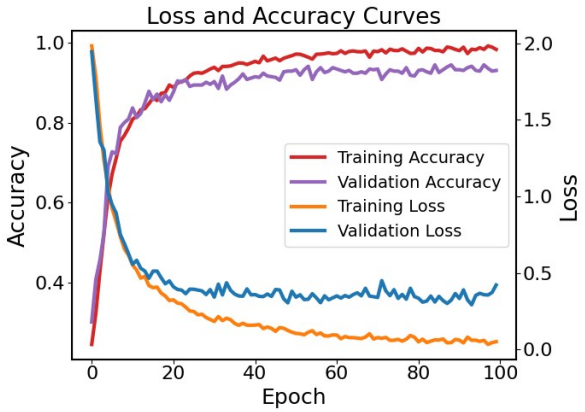

Fig. 5. Model's training and validation plots before modification using SAVEE dataset

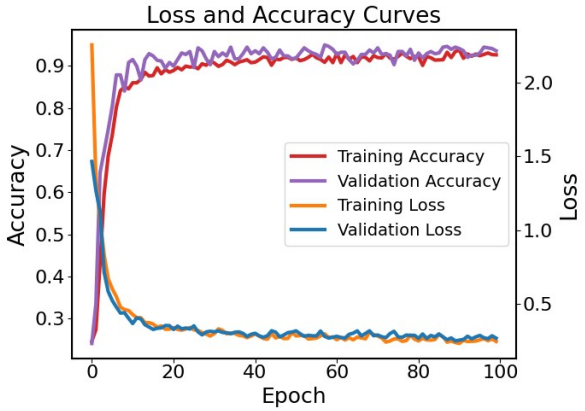

Fig. 6. Model's training and validation plots after modification using SAVEE dataset

unwanted outcome as the model would provide a poor generalization for new data outside of the training set.

To resolve the overfitting problem, the dropout layers were further fine-tuned to find the 'sweet spot' and L1 regularizer was added into the fully connected layer to decay higher weights and penalize the model for larger weights. The percentage of randomized dropout nodes was modified for the last two dropout layers, from 0.3 to 0.5, indicating more nodes were being dropped out before the feature map is forward propagated. The outcome of the training and testing is plotted in Fig. 6, and it shows that the model successfully converged after the modification. Improvement in overfitting after modification infers that high number of filters of last two convolution layers is potentially capturing repeating patterns from the feature map, causing the model to memorize some of the patterns for a particular group.

The proposed model reported a score of 97% and 97% for MEAD, 92% and 92% for RAVDESS, 94% and 92% for SAVEE (after modification), in accuracy and F1-score, respectively. The confusion matrix for each dataset is shown in Fig. 7, Fig 8. and Fig. 9 respectively. The results were further compared with other studies that utilized similar dataset, as summarized in Table I and II. Based on Table I, the proposed model outperformed all the previous literature for RAVDESS to the best of our knowledge, except for the LSTM model by Ryumina and Karpov [24] that considers only the most important distances through thresholding method. From Table II, the proposed model did not outperform majority of the studies with SAVEE dataset in terms of accuracy, but rather at almost same level. But it is still a remarkable performance with

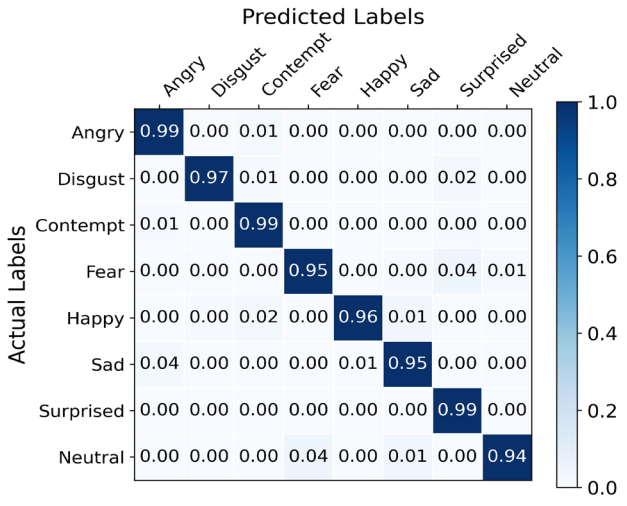

Fig. 7. Confusion matrix of MEAD dataset

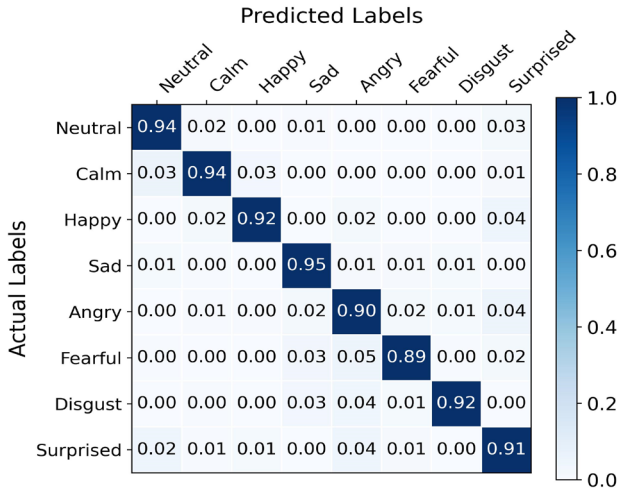

Fig. 8. Confusion matrix of RAVDESS dataset

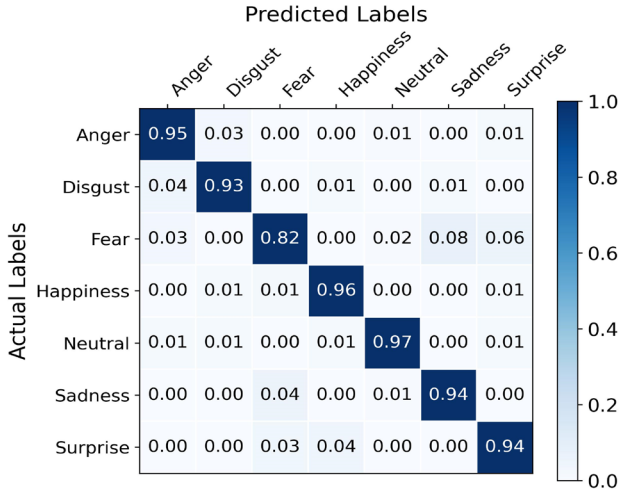

Fig. 9. Confusion matrix of SAVEE dataset (after modification)

a much lower number of parameters (67.14 times lesser compared to [25]) that could contribute to faster computation for lower computation power devices.

IV. Conclusion

In this paper, we proposed a well-generalized CNN model and evaluated its performance by using multiple datasets. As expected, the proposed model drained much lesser computation power as a smaller number of model parameters were recorded, compared to other models from the previous literature. Early results suggested that the application of dropout layer and L1 regularizers in the proposed architecture is one of the main factors that prevented the model from overfitting, hence being more generalized and not susceptible to previously unseen data. Future studies might need to reconsider hyperparameter tuning of the convolution layers to achieve an even better performance.

TABLE I. Comparison of Existing Model with RAVDESS Dataset

Author	Model	Accuracy (%)	F1-score (%)	Parameter #
Ayadi and Lachiri [22]	ResNet50	57.73	-	25,636,712
He *et al.* [23]	AlexNet	79.74	-	62,378,344
Ryumina and Karpov [24]	LSTM (seq. length 122)	98.90	-	63,178,760
Proposed	**Proposed 2D CNN**	**92.00**	**92.00**	**542,136**

LSTM = long short-term memory

TABLE II. Comparison of Existing Model with SAVEE Dataset

Author	Model	Accuracy (%)	F1-score (%)	Parameter #
Sharafi *et al.* [25]	DTN+DSN +CNN	93.75	-	~36.4M
Ngoc *et al.* [26]	VGG-FACE+ LSTM	84.37	-	~134M
Avots *et al.* [27]	AlexNet	94.33	-	62,378,344
Proposed	**Proposed 2D CNN**	**94.00**	**92.00**	**542,136**

DTN = deep temporal network; DSN = deep spatial network; VGG = visual geometry group

References

[1] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, "Facial Expression Recognition with Convolutional Neural Networks via a New Face Cropping and Rotation Strategy," *The Visual Computer*, vol. 36, no. 2, pp. 391–404, Jan. 2019, doi: 10.1007/s00371-019-01627-4.

[2] B. Li and D. Lima, "Facial Expression Recognition via ResNet-50," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 57–64, Jun. 2021, doi: 10.1016/j.ijcce.2021.02.002.

[3] 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, Apr. 2021, doi: 10.3390/electronics10091036.

[4] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network," *Sensors*, vol. 21, no. 9, p. 3046, Apr. 2021, doi: 10.3390/s21093046.

[5] N. Mehendale, "Facial Emotion Recognition Using Convolutional Neural Networks (FERC)," *SN Applied Sciences*, vol. 2, no. 3, Feb. 2020, doi: 10.1007/s42452-020-2234-1.

[6] D. Jeong, B.-G. Kim, and S.-Y. Dong, "Deep Joint Spatiotemporal Network (DJSTN) for Efficient Facial Expression Recognition," *Sensors*, vol. 20, no. 7, p. 1936, Mar. 2020, doi: 10.3390/s20071936.

[7] J. D. Lee and A. Wong, "TimeConvNets: a Deep Time Windowed Convolution Neural Network Design for Real-time Video Facial Expression Recognition," *2020 17th Conference on Computer and Robot Vision (CRV)*, May 2020, doi: 10.1109/crv50864.2020.00010.

[8] Y. Xue, "Comparison of conventional and lightweight convolutional neural networks for Image Classification," *Highlights in Science, Engineering and Technology*, vol. 38, pp. 988–993, 2023. doi:10.54097/hset.v38i.5986

[9] J. Yang, L. Zhang, X. Tang, and M. Han, “CodnNet: A Lightweight CNN architecture for detection of covid-19 infection,” *Applied Soft Computing*, vol. 130, p. 109656, 2022. doi:10.1016/j.asoc.2022.109656

[10] D. Yu, Q. Xu, H. Guo, C. Zhao, Y. Lin, and D. Li, “An efficient and lightweight convolutional neural network for Remote Sensing Image Scene Classification,” *Sensors*, vol. 20, no. 7, p. 1999, 2020. doi:10.3390/s20071999

[11] N. Singh, V. K. Tewari, P. K. Biswas, and L. K. Dhruw, “Lightweight convolutional neural network models for semantic segmentation of in-field cotton bolls,” *Artificial Intelligence in Agriculture*, vol. 8, pp. 1–19, 2023. doi:10.1016/j.aiia.2023.03.001

[12] M. Bodini, “A Review of Facial Landmark Extraction in 2D Images and Videos Using Deep Learning,” *Big Data and Cognitive Computing*, vol. 3, no. 1, p. 14, Feb. 2019, doi: 10.3390/bdcc3010014.

[13] T. Baltrusaitis, P. Robinson, and L.-P. Morency, “OpenFace: an Open Source Facial Behavior Analysis Toolkit,” *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2016, doi: 10.1109/wacv.2016.7477553.

[14] D. E. King, “Dlib-ml: a Machine Learning Toolkit,” *J. Mach. Learn. Res.*, vol. 10, 2009, doi: 10.5555/1577069.1755843.

[15] I. Culjak, D. Abram, T. Pribanic, H. Dzapo, and M. Cifrek, “A Brief Introduction to OpenCV,” in *2012 Proceedings of the 35th International Convention MIPRO*, Opatija, Croatia, 2012.

[16] M. P. Jacob, “Comparison of Popular Face Detection and Recognition Techniques,” *International Research Journal of Modernization in Engineering Technology and Science*, vol. 3, no. 4, pp. 1304–1313, Apr. 2021.

[17] N. Boyko, O. Basystiuk, and N. Shakhovska, “Performance Evaluation and Comparison of Software for Face Recognition, Based on Dlib and Opencv Library,” in *Proc of IEEE Second International Conference on Data Sream Mining & Processing*, Lviv, Ukraine, Aug. 2018, doi: 10.1109/DSMP.2018.847856.

[18] K. Wang *et al.*, “MEAD: a Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation,” *Computer Vision - ECCV 2020*, pp. 700–717, Jan. 2020, doi: 10.1007/978-3-030-58589-1_42.

[19] S. R. Livingstone 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,” *PLOS ONE*, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.

[20] S. Haq and P. J. Jackson, "Multimodal emotion recognition," in *Machine audition: principles, algorithms and systems*: IGI Global, 2011, pp. 398-423.

[21] A. I. Middya, B. Nag, and S. Roy, “Deep Learning Based Multimodal Emotion Recognition Using model-level Fusion of Audio–visual Modalities,” *Knowledge-Based Systems*, vol. 244, p. 108580, May 2022, doi: 10.1016/j.knosys.2022.108580.

[22] S. Ayadi and Z. Lachiri, “Deep Neural Network for Visual Emotion Recognition Based on ResNet50 Using Song-Speech Characteristics,” in *Proc of International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, Hammamet, Tunisia, Mar. 2022. doi: 10.1109/IC_ASET53395.2022.9765898.

[23] Z. He, T. Jin, A. Basu, J. J. Soraghan, Gaetano Di Caterina, and Lykourgos Petropoulakis, “Human Emotion Recognition in Video Using Subtraction Pre-Processing,” *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, Feb. 2019, doi: 10.1145/3318299.3318321.

[24] E. Ryumina and A. Karpov, “Facial Expression Recognition Using Distance Importance Scores between Facial Landmarks,” *Proceedings of the 30th International Conference on Computer Graphics and Machine Vision (GraphiCon 2020). Part 2*, pp. paper32-1paper32-10, Dec. 2020, doi: 10.51130/graphicon-2020-2-3-32.

[25] M. Sharafi, M. Yazdchi, R. Rasti, and F. Nasimi, “A Novel spatio-temporal Convolutional Neural Framework for Multimodal Emotion Recognition,” *Biomedical Signal Processing and Control*, vol. 78, p. 103970, Sep. 2022, doi: 10.1016/j.bspc.2022.103970.

[26] L. B. Ngoc, H.-J. Yang, H. X. Nguyen, S.-H. Kim, G.-S. Lee, and I. S. Na, “Deep Neural network-based Fusion Model for Emotion Recognition Using Visual Data,” *The Journal of Supercomputing*, vol. 77, no. 10, pp. 10773–10790, Mar. 2021, doi: 10.1007/s11227-021-03690-y.

[27] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, “Audiovisual Emotion Recognition in Wild,” *Machine Vision and Applications*, vol. 30, no. 5, Jul. 2018, doi: 10.1007/s00138-018-0960-9