**A Comparative Analysis of Facial Emotion Recognition Models**

Ankita Jaswal

Dept. of Computer Engineering

San Jose State University

San Jose, U.S.A.

ankita.jaswal@sjsu.edu

Jiajun Dai

Dept. of Computer Engineering

San Jose State University

San Jose, U.S.A.

jiajun.dai@sjsu.edu

Akhilandeswari Battineni

Dept. of Computer Engineering

San Jose State University

San Jose, U.S.A.

akhilandeswari.battineni@sjsu.edu

***Abstract—*In today’s society, human computer interaction (HCI) is as prevalent as ever. Recent applications of HCI have drawn their attention towards Facial Emotion Recognition (FER) in various domains to improve the user experience when dealing with HCI. While most approaches focus on providing the most accurate solution, they fail to optimize the performance of their models. Considering the current need for fast performing models in real-time applications such as mobile devices, it is crucial to not only focus on the enhancement of this area of research, but also find lightweight solutions with high performance. This paper provides a comparison of recent state-of-the-art (SOTA) solutions that focus on this problem. Models such as: EfficientNet, MobileNetV2, and VGG16 with transfer learning are implemented and evaluated using a custom dataset. From the three models MobileNetV2 yields 2.5M training parameters and 60.70% accuracy resulting in our proposed approach when seeking out a lightweight yet optimized solution for FER in real-time applications.**

***Index Terms –* Facial Emotion Recognition (FER), HCI, EfficientNet, MobileNetV2, VGG16**

1. INTRODUCTION

In recent years, as the internet services continue to grow rapidly, the area of facial emotion recognition (FER) has gained massive attention in various domains. The demand for fast performing solutions to automatically detect facial expressions in the fields of image recognition, computer vision, and HCI has primarily taken a surge [1]. The reason being able to analyze how a user feels when interacting with a device can help industries provide more efficient solutions to their customers by focusing on their emotional activity. This method of communication can further improve applications and internet services provided to customers by different domains. For example, users in the scene of online learning platforms may not be learning as efficiently, being able to analyze the emotions of the user can help online learning platforms to provide personalized solutions according to user needs which in terms improves their overall ratings and user satisfaction [1]. Hence why there is a recent growing interest in the field of FER.

In order to perform the task of FER, the system must be able to get faces frame-by-frame, extract important facial features and perform pre-processing steps, and classify the facial emotion using an efficient and lightweight model. This helps provide the current emotional state of a person by classifying the type of expression a person is expressing.

In this paper, we primarily focus on the problem of finding one of the most optimal solutions for FER that is not only lightweight, but also high in performance in terms of accuracy. For this case, we provide a comparison of three popular approaches for facial emotion recognition: MobileNetV2, EfficientNet, and transfer learning with VGG16. All three approaches are implemented and evaluated using our custom dataset which is a collection of three open source FER datasets. This includes FER2013 [2], Emotic [3], and MH-FED [4]. These datasets are processed and a total of 31K good samples are extracted for seven base face expressions: angry, disgust, surprise, happy, sad, neutral, and fear.

The structure of this paper is as follows. In section II, we will discuss related works in the context of emotion recognition applications. Section III will cover our proposed methodology. Section IV includes a discussion of our proposed application evaluation results, real-time application results, and limitations. At last, section V will conclude our study and provide some direction to future researchers in the field of facial emotion recognition.

1. RELATED WORKS

Emotion recognition has been researched in various computer vision applications. These applications include emotion recognition in context, video, audio, and facial emotion recognition. Primarily researchers focus on the analysis of facial expression by predicting facial emotion [3]. The process used for this specific task is getting localized features of the face in a frame, action units [3]. These features are often bounded by a bounding box and landmarked by facial landmarks. Then, these extracted features are used to recognize the emotions in their specified categories [5].

Since this task heavily involves image recognition tasks, recent state-of-the art approaches use convolutional neural networks (CNN) based models for emotion recognition [6]. In A-MobileNet, the authors use the MobileNetV1 architecture with a combination of convolutional block attention modules (CBAM) [12] in between each depth-wise separable block (DSC) to create their own custom model for FER task [7]. In another paper, the authors use DCNN for this task primarily focusing on non-frontal faces [9].

In addition, the performance of a model can depend entirely on the proposed dataset. If the dataset is biased, it will lead to poor performance in real-time. In some cases preprocessing steps can be taken to avoid this measure. In a survey conducted by [6] many state-of-the-art solutions use similar processing techniques. This includes, but not limited to, grayscale conversion, dimensionality reduction, facial localization and adjustment, resizing, and data augmentation.

Another important aspect considered by many researchers for the FER task is feature extraction. This helps in reducing the computational complexity for the task at hand. The authors of [3], mention using a one-dimensional filter CNN based approaches for feature extraction. While another study proposed using MTCNN for extracting facial regions for recognition tasks [8].

1. METHODOLOGY
   1. *Dataset Collection & Preprocessing*

In our paper, we utilized a range of datasets to train and evaluate our facial emotion recognition models. These datasets include:

1. Emotic Dataset: This dataset comprises a wide variety of facial expressions captured in many contexts [3]. It samples many datasets to provide a good representation of data for researching context based facial emotion recognition such as: MSCOCO, Emodb-small, Framedb, and Ade20k.
2. FER2013 Dataset: This dataset is known for consisting of facial expressions mainly for FER tasks. This data is one of the most baselined datasets for this task containing seven base emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral [2].
3. MH-FED: This dataset offers a unique set of samples including micro-expressions which are important for detecting even the most subtle emotional changes in the eyes and mouth regions. This dataset was generated using a GAN based model consisting of 10 META-human characters.

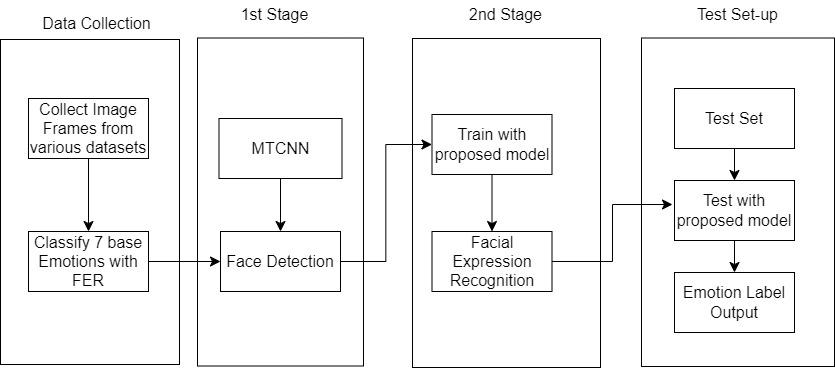


**Fig. 1** emodb\_small sample preprocessed image

Each dataset was carefully selected to ensure that a broad spectrum of representations are considered for facial emotion recognition to reduce the induction of bias in our proposed models.

All data was preprocessed to align with each model’s specific requirements, mainly focusing on facial features enhancements for improving accuracy. To keep a uniform approach across all comparisons. All data was passed through a face extraction module using DNN’s CaffeModel and passed through FER pre-trained model for automatic labeling. Fig.1 shows the process taken and resulting images.

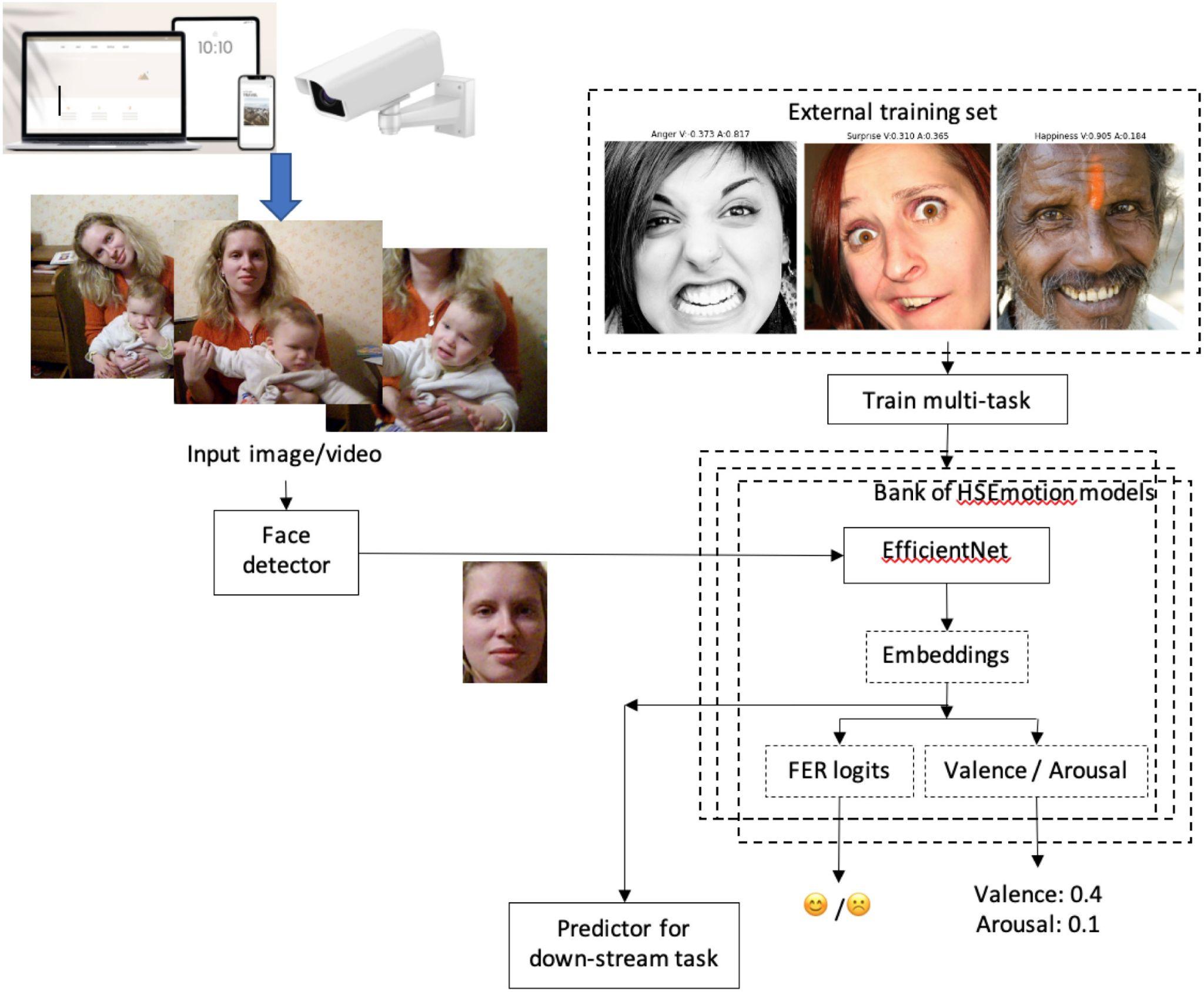
* 1. *System Design Approach*



**Fig. 2** System Design Overview

To keep a uniform application for comparing different models, our system was designed using a multi-stage approach. After Data Collection steps, we feed the data to two stages. In the first stage, Face Detection is completed using MTCNN [8]. This results in storing the region of interest being facial landmarks as key-points: right eye, left eye, right mouth, left mouth, and nose as well as storing the bounding box coordinates. This information is fed into the second stage where our proposed models are trained for facial expression recognition tasks. At last, the trained models are evaluated using our test set based on the output of emotion label probabilities. A general overview of the proposed design is illustrated in Fig. 2.

* 1. *EfficientNet Architecture*



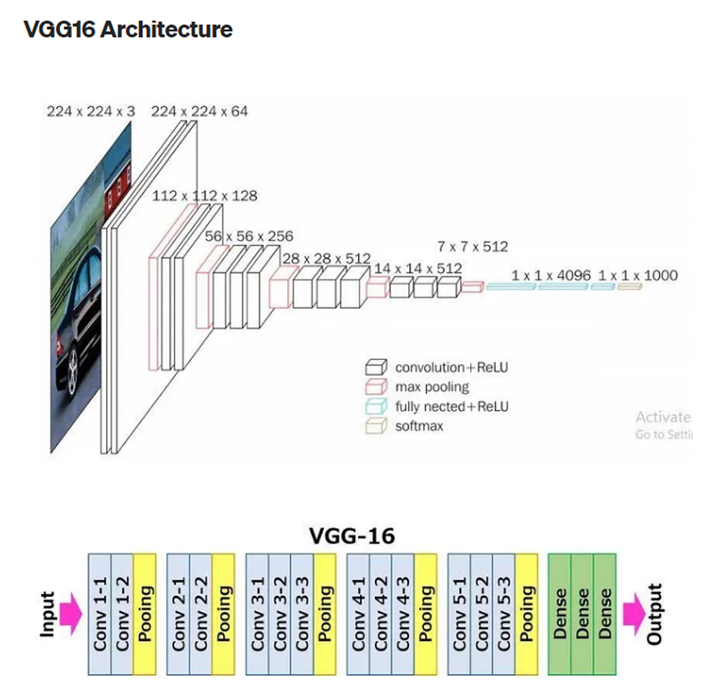
**Fig. 3** HSEmotion Pipeline. Image Source [10]

The first model used in the set of comparisons of our paper is EfficientNet. A lightweight CNN approach which offers improved performance and scalability with much fewer trainable parameters than traditional CNNs. For the case of our choice, we chose to utilize a state-of-the-art method provided by HSEmotion PyPi library [10]. The authors of this paper have curated a set of EfficientNet models oriented for multi-task training. Of the trained tasks, it includes FER logits, a way to perform facial emotion recognition and output the 7 base emotion probabilities noticed in the original FER2013 dataset. The list of available models is as follows:

1. Enet\_b0\_8\_best\_afew
2. Enet\_b0\_8\_best\_vgaf
3. Enet\_b0\_8\_va\_mtl
4. Enet\_v2\_7
5. Enet\_v2\_8

All of their models are made for high-speed emotion recognition to use in low powered devices like mobile phones. They are pre-trained for face identification tasks using VGGFace2 dataset and fine-tuned using the EfficientNet-B0 timm library implementation for emotion recognition [11]. Fig. 3 provides an insight into the entire end-to-end pipeline for using any one of the HSEmotion models.

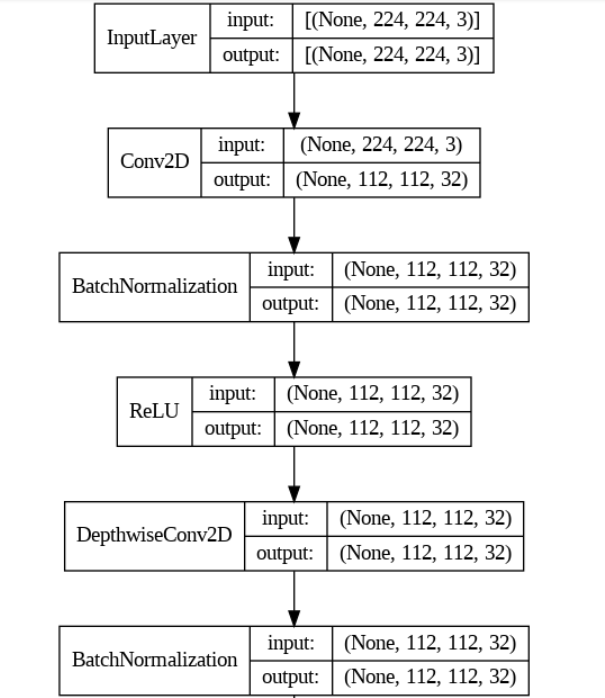
* 1. *FER VGG16 Architecture*



**Fig. 4** VGG16 Architecture. Image Source [13]

The second model used for comparison is VGG16. A 16-layer network architecture with weights trained on the ImageNet dataset. For this project, we chose to use the trained VGG16 model provided by Keras api for transfer learning and fine tuning. To fine tune our model, we first froze the first fourteen layers (ten convolutional layers and four max-pooling layers) of the model and only train the last convolutional block with our own dataset. Since our dataset is imbalanced, meaning the number of images in “fear” and “disgust” is in the minority. We used image augmentation as a technique for our VGG model to increase the variation of our training dataset by applying horizontal and vertical flip of the original images.

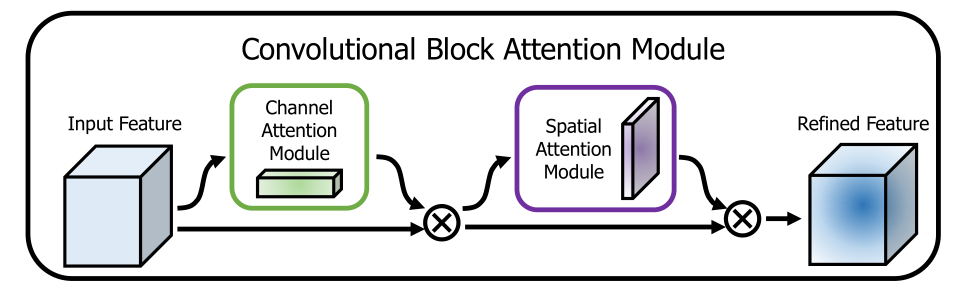
* 1. *MobileNet Architecture*

**

**Fig. 5** Mobilenetv2 Architecture

Third Model used for comparison is MobileNetV2 is a convolutional neural network (CNN) that is designed to be highly efficient and accurate for image classification tasks. It achieves this efficiency through the use of depth wise separable convolutions, which reduce the number of parameters and computational cost while maintaining good performance. MobileNetV2 consists of a stem layer, multiple inverted residual blocks, and a classification layer. The stem layer is a standard convolutional layer that processes the input image. The inverted residual blocks are a series of layers that perform feature extraction and downsampling. Each block consists of a convolution,Depthwise convolution,ReLu The classification layer is a fully connected layer that produces the final output of the network

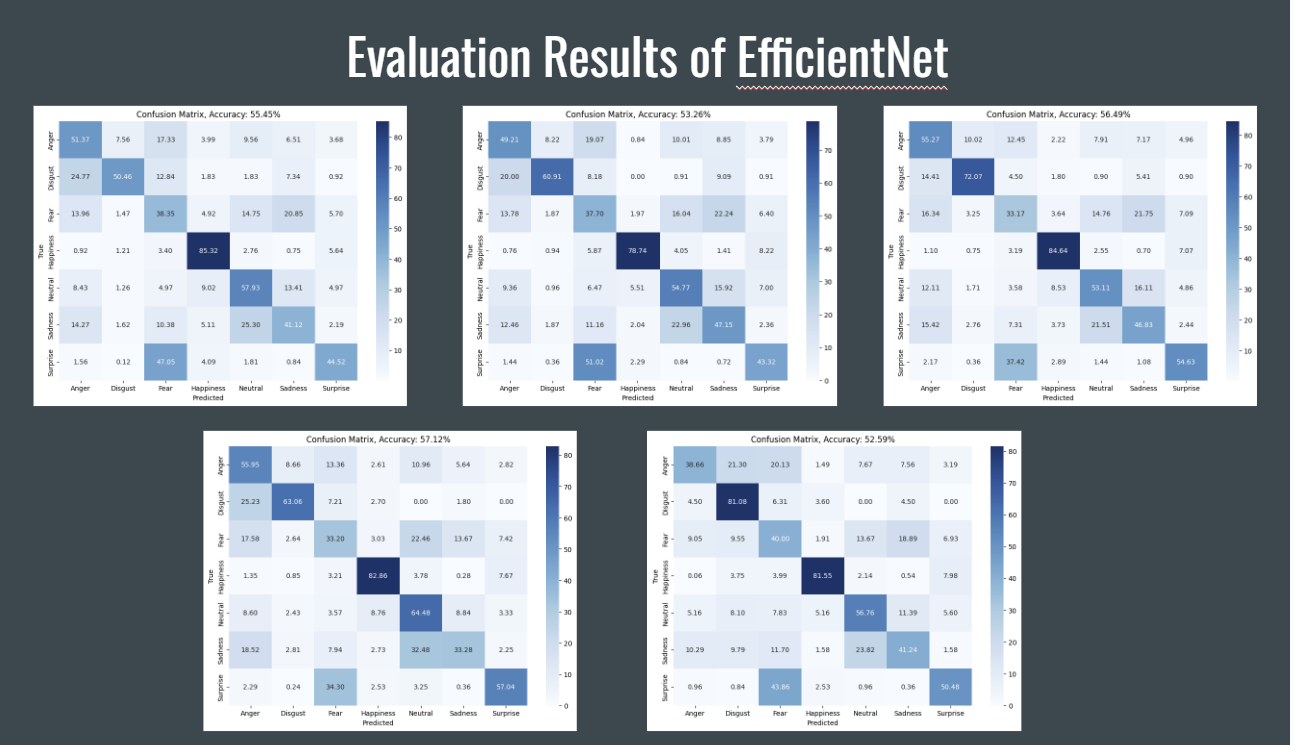
* 1. *MobileNetV2\_CBAM* Architecture



**Fig. 6** Mobilenetv2\_CBAM Architecture

We Applied CBAM model in the DSC block of MobileNetV2 to increase the accuracy of the model . CBAM helps to Uses Channel Attention Module and Spatial Attention Module for retaining information, normally lost by applying ReLU

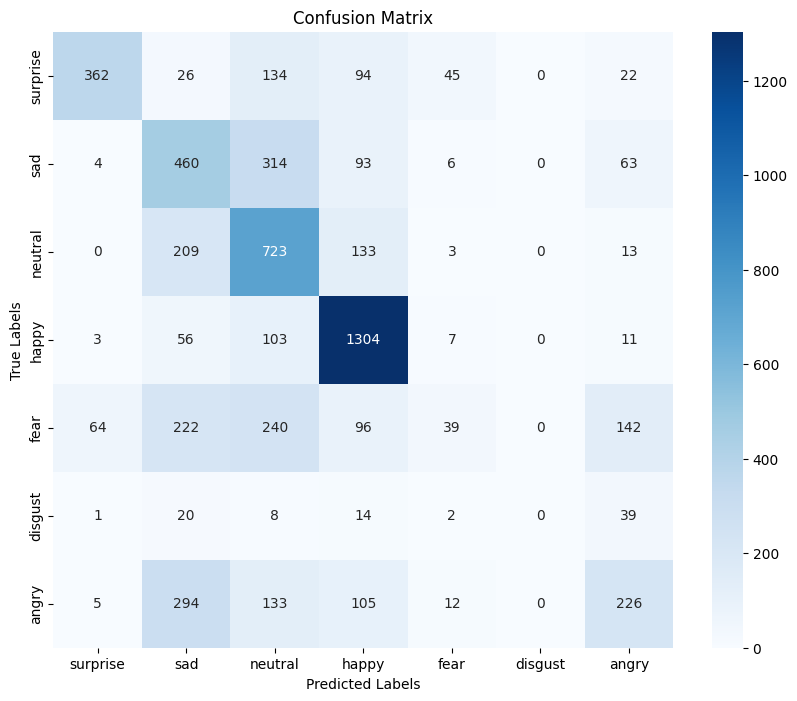
1. RESULTS
   1. *Training and Evaluation Results*

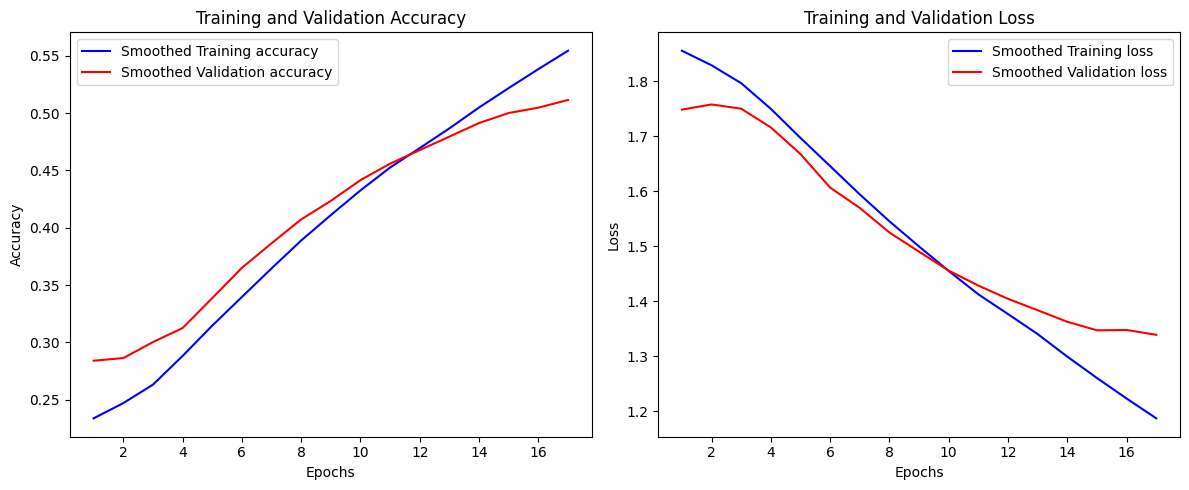


**Fig. [7]** EfficientNet evaluation results for all pretrained models. The bottom left image is best performing of all provided models.

1. HSEmotion EfficientNet

Since HSEmotion EfficientNet models are pre-trained for our task, we will be providing the results gathered after evaluating each on our test set. As shown in Fig. [7], Enet\_v2\_7 resulted with the best accuracy at 57.12% and the probabilities of actual emotion label v. predicted labels are shown in each of their confusion matrices in the same order as the list shown in Part C of our methodology.

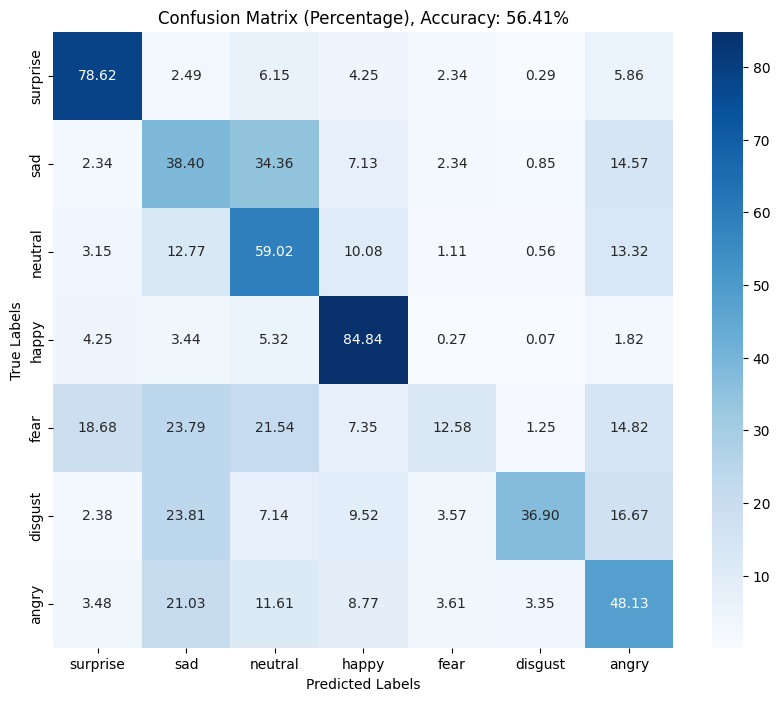
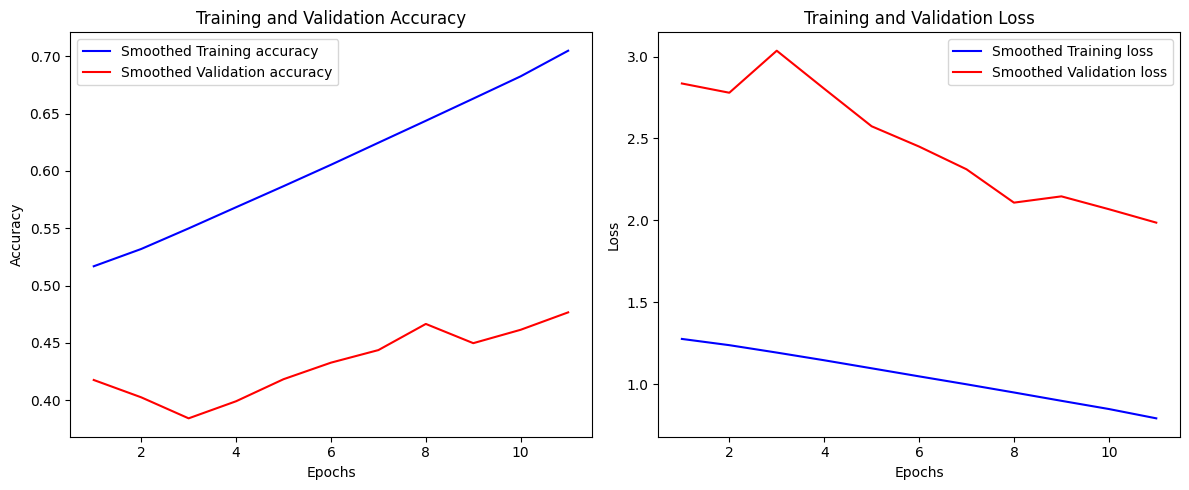




**Fig. [8]** VGG16 evaluation results for the pretrained models.

2. VGG16 finetune

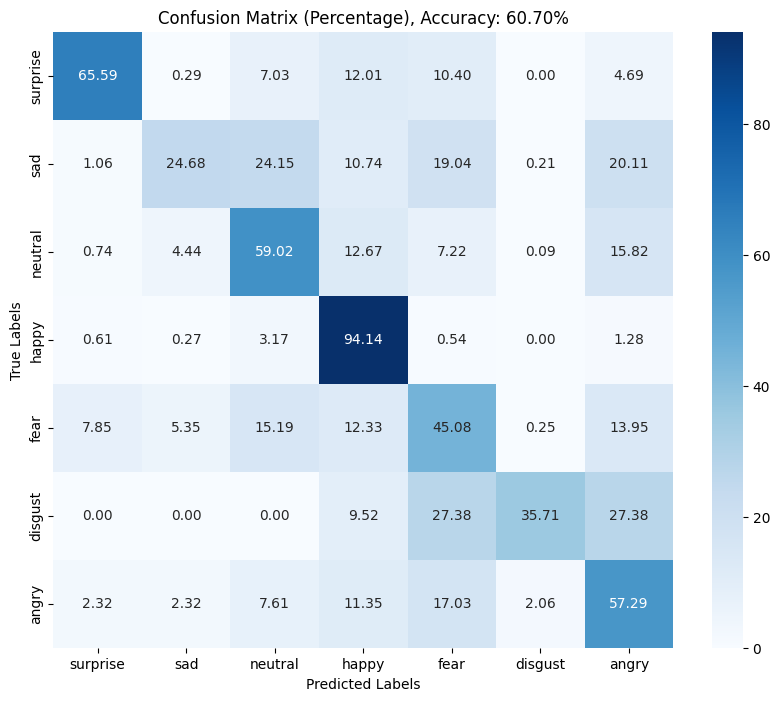
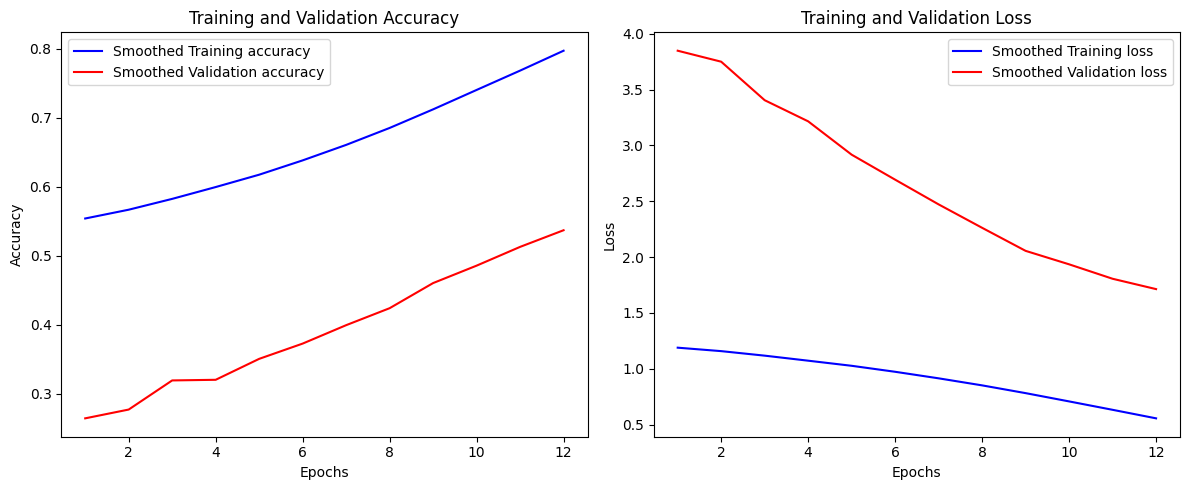
As shown in Fig. [8], our VGG16 training stopped at epoch 17 with the best accuracy at 58.5% and F1 score as 0.53. It still performs poorly on the minority emotions after we applied image flipping onto the dataset.



**Fig. [9]** MobileNetV2 evaluation results for the pretrained models.

3. MobileNetV2

As shown in Fig. [9], our mobilenetV2 model accuracy is 56.41%.



**Fig. [10]** MobileNetV2 \_CBAM evaluation results for the pretrained models.

4. MobileNetV2\_CBAM

As shown in Fig. [10], our mobilenetV2\_CBAM model accuracy is increased from mobilenetv2 model from 56.41% to 60.70%

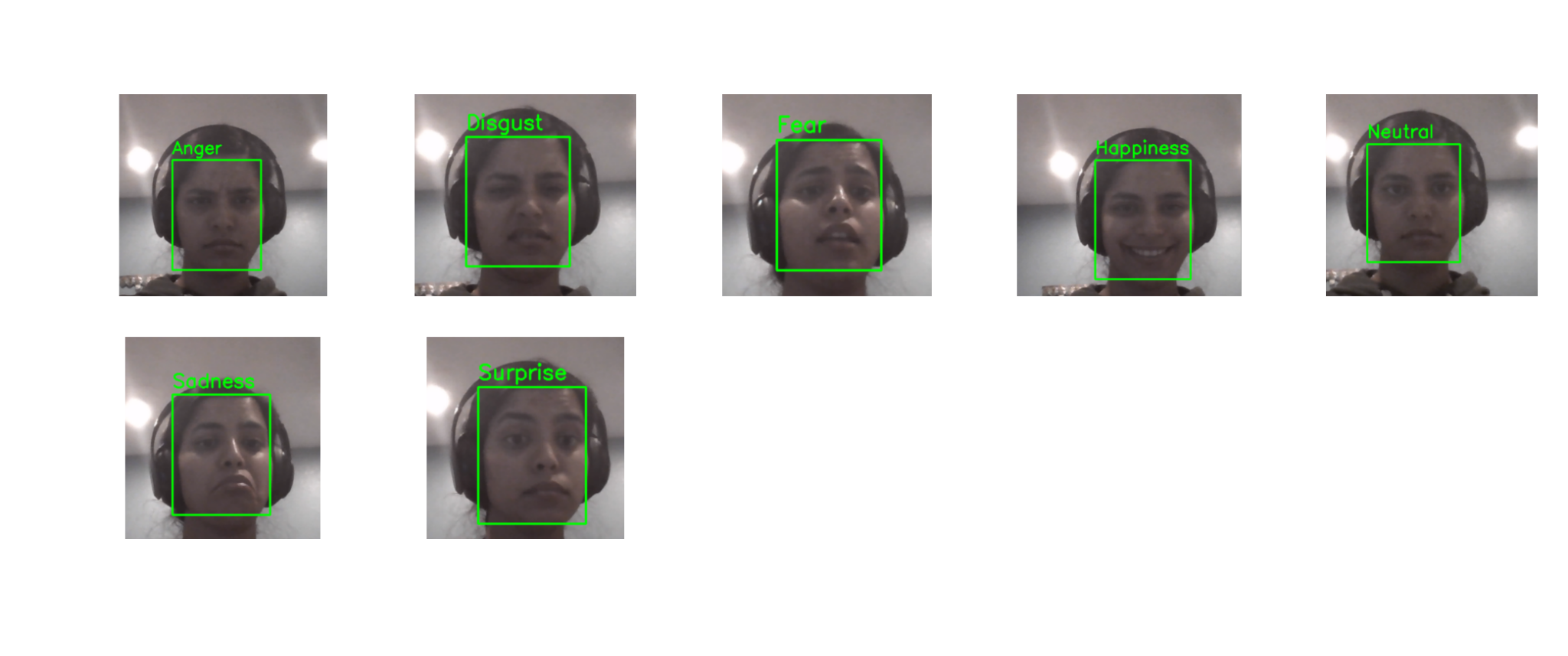
**TABLE I**

Comparison Results of all models obtained on test set

| **Model** | **Accuracy (%)** |
| --- | --- |
| EfficientNet | 57.14 |
| MobileNetV2 | 56.41 |
| MobileNetV2\_CBAM | 60.70 |
| VGG16\_finetuned | 58.50 |

* 1. *Real-Time Test Results*

After gathering the evaluation results from testing each one of our proposed models, it is evident that MobileNetV2 implementation with the CBAM architecture results in best performance and least amount of trainable parameters as demonstrated in Table I. Hence, for our real-time testing we are using HaarCascade and this saved model. Fig.[7] demonstrates the results for each emotion category of the 7 which the model was trained on.



**Fig.[11]** Real-Time FER results with MobileNetV2\_CBAM

* 1. *Limitations*

Some limitations faced during the process of our comparisons for the model is fighting model bias. Our dataset was significantly imbalanced in the categories for “Sadness”, “Fear”, and “Anger” . This resulted in micro-emotions such as only using the eye region or only using the mouth region for emotion recognition. If either of the regions of interest information was missing, the models led to misclassifications.

As a result, the models were retrained using data augmentation techniques, resizing of the images, turning all images to grayscale, and shuffling the train data prior to splitting.

1. CONCLUSION and FUTURE WORKS
   1. *Conclusion*

Although our real-time facial emotion recognition model is imperfect due to the imbalance of image data between majority and minority emotions, we were able to achieve a satisfying result with our best performing model - the MobileNetV2 with CBAM. Our real time testing model was able to identify all facial emotions, except that sometimes we need to exaggerate the minority emotions (sad, fear, anger, disgust) for the model to recognize.

* 1. *Future Works*

In order to further improve the accuracy of our facial emotion recognition model in the future, we will apply different techniques (such as over-sampling, data augmentation) to reduce the imbalance of our dataset. In addition, we will also try to train different models, including but not limited to CNNs, to find the model that has the best performance for our real-time facial emotion recognition application.

1. CONTRIBUTIONS

**Ankita Jaswal** - Data collection, preprocessing, EfficientNet training/evaluation, MobileNetV2 CBAM training/evaluation, real-time testing using haar-cascade and best performing model.

**Jiajun Dai** - VGG16 transfer learning with fine-tuning training/evaluation

**Akhilandeswari Battineni** - MobileNetV2 model training /evaluation

**All** - Project Report and Presentation

**Google Drive Link**: https://drive.google.com/drive/folders/0AP5R7FWbkUOEUk9PVA

REFERENCES

[1] A.T. Lopes, E. De Aguiar, T. Oliveira-Santos, A facial expression recognition system using convolutional networks, in: 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images, IEEE, 2015, pp. 273–280.

[2] Y. Fan, J. Lam, V. Li, Multi-region ensemble convolutional neural network, Springer International Publishing, 2018.

[3] R. Kosti, J.M. Álvarez, A. Recasens and A. Lapedriza, "Context based emotion recognition using emotic dataset", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2019.

[4] J. R. Siddiqui, "FExGAN-Meta: Facial Expression Generation with Meta-Humans," ArXiv e-prints, [Online]. Available: https://arxiv.org/abs/2203.05975, 2022, doi: 10.31219/osf.io/ygdrt.

[5] P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion,” J. Personality Social Psychology, vol. 17, no. 2, 1971, Art. no. 124.

[6] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski, “A survey on facial emotion recognition techniques: A state-of-the-art literature review.”, Inf. Sci. 582, C, 2022, 593–617. https://doi.org/10.1016/j.ins.2021.10.005

[7] Y. Nan, J. Ju, Q. Hua, H. Zhang, and B. Wang, ‘A-MobileNet: An approach of facial expression recognition’, Alexandria Engineering Journal, vol. 61, no. 6, pp. 4435–4444, 2022.

[8] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," in IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[9] Z. Yu, C. Zhang, Image based static facial expression recognition with multiple deep network learning, in: Proceedings of the 2015 ACM on international conference on multimodal interaction, 2015, pp. 435–442.

[10] A. V. Savchenko, "HSEmotion: High-speed emotion recognition library," Software Impacts, vol. 14, p. 100433, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665963822001178. [Accessed: Dec. 2, 2023]. DOI: 10.1016/j.simpa.2022.100433.

[11] R. Wightman, ‘PyTorch Image Models’, GitHub repository. GitHub, 2019.

[12] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[J]. 2018. ECCV2018

[13] R. G, “Everything you need to know about VGG16,” Medium, https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918 (accessed Dec. 4, 2023).