# Image-Based Facial Emotion Recognition

Abu Bakar Siddik Hasan
ID-u3237372
University of Canberra
ACT, AUSTRALIA

Somya Shukla
ID-u3238876
University of Canberra
ACT, AUSTRALIA

**Abstract**

Accurate Facial Emotion Recognition (FER) is an essential component of human-computer interaction and emotional understanding, impacting diverse fields including safety, healthcare, and human-machine interfaces [1]. The core challenge in FER lies in precisely identifying human emotions amidst the intricate diversity of facial expressions and their varying presentations. In our study, we tackle this central issue by conducting experiments using a 5-layer custom CNN and four pre-trained models (ResNet50, VGG16, MobileNet, and Inception3) on the FER-2013 dataset, categorizing facial expressions into seven emotions (Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral). Notably, our best-performing model, MobileNet, achieved a training loss of 1.5114 and a training accuracy of 0.4194. In the validation set, it recorded a loss of 1.4696 and an accuracy of 44%. Furthermore, on the test set, this model demonstrated a noteworthy performance, achieving an accuracy of approximately 43%. This research contributes to resolving the persistent problem of accurately identifying human emotions from facial expressions in real-world conditions, thus advancing the field of Facial Emotion Recognition.

## 1 Introduciton

Imagine a computer that can understand how you feel just by looking at your face. This is what we call Facial Emotion Recognition, or FER. It's a powerful tool that can improve how we interact with computers and how we understand emotions [2]. FER has important applications in keeping people safe, managing healthcare, and making machines work better with us. However, FER is a tough problem. Faces can look very different, and sometimes, it's hard to tell one emotion from another. In our study, we're trying to make this process better.We did experiments using different kinds of computer models and a special dataset with pictures of people's faces showing seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral expression. We trained our models to get really good at this task. Our best model, called MobileNet, did especially well. It learned to recognize these emotions very accurately. We trained it to be good at this by showing it lots of pictures. When we tested it, we could tell how people were feeling in pictures with about 43 percent accuracy.

Our research aims to help solve the challenge of recognizing emotions from people's faces in real-life situations. This can make FER technology even more useful and helpful in our daily lives. It's all about making technology work better with us and understanding our feelings.

# 2 Motivation

The motivation for this project is to develop a robust and accurate FER system that can be used to improve our understanding of emotions and how we use this knowledge to improve our lives. We believe that FER systems have the potential to make a significant impact on a variety of fields, including psychology, technology, and healthcare [3] [4]. For example, FER systems could be used to develop new tools for diagnosing and treating mental disorders, to develop more engaging and interactive social robots, and to help people with communication disorders communicate more effectively. We are also motivated by the challenge of developing an FER system that can achieve state-of-the-art accuracy. FER is a challenging task due to the high variability of human facial expressions and the complex relationship between facial expressions and emotions. However, we believe that recent advances in deep learning make it possible to develop FER systems that can achieve unprecedented levels of accuracy. We are excited to contribute to the field of FER and to develop FER systems that can make a real difference in the world.

# 3 Literature review

Facial emotion recognition has gained significant attention in the fields of computer vision and machine learning. Numerous studies have examined various approaches to address this complex problem. Deep learning algorithms have been used by many researchers to obtain great accuracy in facial emotion recognition.

Durga and Rajesh (2022) [5] developed a 2D-ResNet model using the JAFFE dataset for facial recognition. Their method, compatible with Flask, MySQL, Fastai, and JSON, achieved remarkable metrics, including 99.3% accuracy, 99.12% recall, 98% F1 score, and 99.16% sensitivity, with just 5 failures in 100,000 tests. In the year 2022, Moravčík and Basterrech [6] discussed the importance of understanding human emotions while interacting with machines. They looked at the effects of Transfer Learning (TL) for Facial Emotion Recognition (FER) on a Convolutional Neural Network using a VGG-type architecture. They evaluated the CK+ dataset after applying TL to improve model accuracy and generalization using the FER2013 dataset. This work demonstrates how TL can enhance deep learning models' ability to recognize basic emotions.

A system for recognizing emotions was created by Sujanaa, Palanivel, and Balasubramanian (2021) [7], with a focus on happy, normal, and surprised feelings. Their method used mouth pictures that were taken from video frames and segmented at a rate of 20 frames per second using a Haar-based cascade classifier. To extract features and combine them into a single histogram, edge and local information were captured using the Histogram of Oriented Gradients (HOG) and Local Binary Pattern (LBP). In addition, unique points were extracted using Scale-Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF). These texture features were used to train one-dimensional Convolutional Neural Network (1D-CNN) and Support Vector Machine (SVM) models. Impressive accuracy was shown in the experimental findings, where SVM achieved 97.44% and 1D-CNN achieved 98.51%.To recognize facial expressions, Wang et al. (2019) [8] presented a hybrid feature technique that combines deep-learning features taken from a CNN model with SIFT features. When employed for classification, Support Vector Machines (SVM) produced better results than the most advanced CNN techniques. To explore deep neural networks and regression models for object detection and posture estimation, Hara completed her doctoral dissertation at the University of Maryland in 2016 [9]. Her research offers important new insights into computer vision and machine learning. Using the FER-2013 dataset, Kusuma, Jonathan,

and Lim (2020) [10] developed an emotion identification system that improved the VGG-16 CNN. Outperforming many standalone-based models, the model classified seven unique emotions with an accuracy of 69.40%. For facial recognition, Dubey, and Jain (2020) [11] used a transfer learning model based on VGG16. Their method outperformed previous methods with accuracy scores of 94.8% and 93.7% on the CK+ and JAFFE datasets. Chowdary, Nguyen, and Hemanth (2021) [12] combined transfer learning with Resnet50, VGG19, Inception V3, and MobileNet to concentrate on deep learning-based facial emotion recognition. By adding new fully connected layers specifically designed for their goal, they modified the model and were able to achieve an average accuracy of 96% for emotion identification on the CK+ database. Using VGG-16 and optimized CNN models, Vatcharaphrueksadee, Viboonpanich, Sakul-ang, and Maliyaem (2020) [13] investigated emotion categorization. Their goal was to achieve at least 65% accuracy in identifying five fundamental emotions, with the goal of improving the accuracy and training efficiency of the most advanced model. A very accurate Facial Emotion Recognition (FER) system employing Transfer Learning (TL) with a Very Deep Convolutional Neural Network (DCNN) was introduced by Akhand, Roy, Siddique, Kamal, and Shimamura (2021) [14]. Using the KDEF and JAFFE datasets, they optimized pre-trained DCNN models (VGG-16, VGG-19, ResNet, Inception-v3, and DenseNet-161) for FER. Their method outperformed other FER systems in terms of accuracy, with DenseNet-161 scoring 96.51% on KDEF and 99.52

Xia, Xu, and Nan (2017) [15] investigated the field of facial expression recognition, which has a variety of uses in machine vision, pattern identification, and human-machine interaction. They retrained the Extended Cohn-Kanade dataset using transfer learning techniques after adopting the Inception-v3 model on the TensorFlow framework. This approach greatly decreased training time without sacrificing recognition accuracy. Facial Expression Recognition (FER) systems were the subject of research by Alam, Kartowisastro, and Wicaksono (2022) [16] with applications in the fields of law enforcement, marketing, healthcare, and education. Using transfer learning, they used Deep Convolutional Neural Networks (EfficientNet family). The method improved the performance of FER models with impressive accuracy, especially when using the EfficientNet-B0 architecture, which yielded 99.57% accuracy on CK+ and 100% accuracy on JAFFE datasets in the test set. The topics of generalization and robustness in face emotion identification were discussed by Li and Lima (2021) [17]. In comparison to conventional models, they introduced a feature extraction technique that combined a convolutional neural network with ResNet-50 to achieve higher face emotion identification performance. Vepuri (2021) [18] improved facial emotion recognition with CNNs and preprocessing techniques, with an accuracy of 69.46%. An ensemble accuracy of 76.01% was achieved using transfer learning using pre-trained models such as Resnet-50, Senet-50, VGG16, and FaceNet.

The review of the research concludes by showcasing a variety of approaches in the field of facial expression recognition and emphasizing the important contributions made by different models and strategies in improving performance and accuracy in this area.

# 4 Data Description

The 2013 Facial Expression Recognition dataset, also known as FER2013, was created and shared on Kaggle. It was first introduced at a machine learning conference in 2013 by Pierre-Luc Carrier and Aaron Courvill [19]. This dataset contains pictures of people's faces, and each face is sorted into different categories based on their emotions. The images in the FER-2013 dataset are black and white and measure 48 pixels by 48 pixels. In total, there

are 35,887 pictures in this dataset, and they represent seven different types of tiny facial expressions. Each expression is given a label, which is a number from 0 to 6.

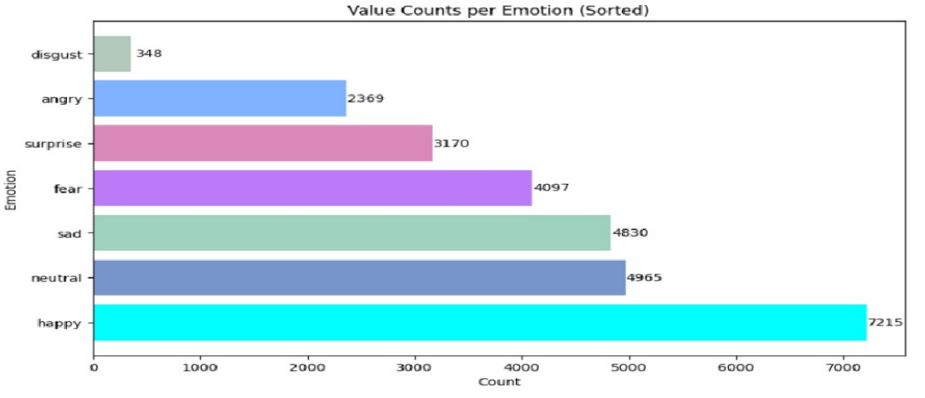


Figure 1: Values count per emotions

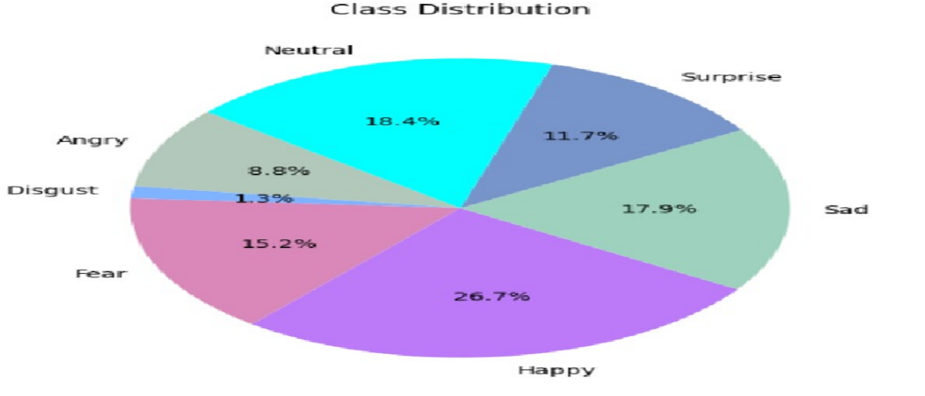


Figure 2: Class Distribution

# 5 Data Preprocessing

Data preprocessing is an essential step in preparing the FER2013 dataset for training a facial emotion recognition model. This process ensures that the model can effectively learn from the data and make accurate predictions. We employ TensorFlow's ImageDataGenerator to augment the data, which includes rescaling the pixel values, allocating a portion of the training data for validation, and applying random shifts, rotations, shearing, and flips to the images. These augmentations significantly enhance the model's capacity to generalize to new data. Additionally, we establish a consistent target image size of 48x48 pixels for all models, except for MobileNet, where the minimum requirement is 75x75 pixels. We also specify a batch size of 128 to ensure uniform processing.

# 6 Model Selection

We evaluated five models, including MobileNet, Custom CNN, InceptionV3, ResNet-50, and VGG16, for facial emotion recognition based on performance metrics.

## 6.1 Custom CNN (Model 1)

The custom CNN OR Model 1, was designed with 2D convolutional layers followed by batch normalization, max-pooling layers, and dropout layers [21]. This model architecture consists of three convolutional layers, two with a 3x3 kernel and one with a 5x5 kernel. It includes two max-pooling layers and dropout layers with a rate of 0.2 to promote generalization. The CNN is followed by three fully connected layers with 512, 256, and 7 neurons in the output layer to predict the emotion classes.

## 6.2 MobileNet (Model 2)

A convolutional neural network architecture called MobileNet was used to extract features from the input photos. In this concept, extra layers for classification are added to the basic MobileNet architecture. The architecture of MobileNet contains many convolutions, which are helpful in lowering computational complexity without compromising performance [20]. To prevent overfitting, we added two dense layers with 1024 and 512 neurons, respectively, and dropout layers with a rate of 0.5 to further refine the model. Ultimately, seven output units of a softmax layer were added to categorize seven distinct emotions.

## 6.3 InceptionV3 (Model 3)

The third model, InceptionV3, is a more complex architecture [22]. The base model, pre-trained on ImageNet, was fine-tuned for our emotion recognition task. We added dense layers to perform the final classification. The architecture includes global average pooling, followed by two dense layers with 1024 and 512 neurons, respectively, and dropout layers with a rate of 0.5. The output layer uses softmax activation to predict one of seven emotion classes.

## 6.4 ResNet-50 (Model 4)

In Model 4, we used the ResNet-50 architecture, a deep neural network [23]. The pre-trained ResNet-50 model was used as a feature extractor. The top layers of ResNet-50 were made non-trainable, and a custom top model was added. The top model includes two dense layers with 1024 and 512 neurons, along with dropout layers with a rate of 0.5. Softmax activation is used in the final layer to classify emotions into multiple classes.

## 6.5 VGG16 (Model 5)

Model 5 is based on the VGG16 architecture, pre-trained on ImageNet [24]. We integrated the VGG16 base model with additional dense layers. The custom top model includes global average pooling, followed by two dense layers with 1024 and 512 neurons, and dropout layers with a rate of 0.5. The output layer employs softmax activation to classify images into seven different emotion categories.

Each of these models was trained and evaluated using a common set of data preprocessing steps and augmentation techniques. After training, we assessed their performance on a testing dataset and monitored key metrics, including accuracy. The use of early stopping helped prevent overfitting during training. The results and performance metrics of these models are discussed in subsequent sections, allowing us to select the most effective model for facial emotion recognition.

# 7 Model Training

With the FER2013 dataset appropriately preprocessed, we proceeded to train our selected models, each designed to recognize facial emotions from the grayscale 48x48 pixel images. Below, we provide a short overview of the training process for each model.

## 7.1 Custom CNN (Model 1)

Our customary convolutional neural network, Model 1, was trained using the Adam optimizer with an initial learning rate of 0.001. As the loss function, categorical cross-entropy was utilized. For best results, early stopping was implemented during the 50 epochs of training. The training dataset performed a similar application of data augmentation techniques. The model was composed of max-pooling, convolutional, batch normalization, and dropout layers to reduce overfitting. To provide multi-class emotion categorization, three completely connected layers were implemented.

## 7.2 MobileNet (Model 2)

The Adam optimizer was used to train Model 2, MobileNet, with categorical cross-entropy serving as the loss function. To avoid overfitting, the training procedure was conducted over 50 epochs with early termination. By using data augmentation techniques to the training dataset, the model's capacity to generalize to previously unseen data was improved. The architecture resulted in a softmax output layer for the classification of seven emotions, which was utilized by two dense layers and dropout layers.

## 7.3 InceptionV3 (Model 3)

Model 3, InceptionV3, featured a fine-tuning approach. The model was pre-trained on ImageNet, and specific top layers were unfrozen for emotion recognition. The training process employed the Adam optimizer with categorical cross-entropy loss. Early stopping was introduced, ensuring model generalization. InceptionV3 included global average pooling and two dense layers, followed by dropout layers. The final classification was executed through a softmax output layer.

## 7.4 ResNet-50 (Model 4)

Model 4 utilized the ResNet-50 architecture, initially pre-trained on ImageNet. The training involved unfreezing the top layers of the base model and adding a custom top model. The training process, employing the Adam optimizer, proceeded for 40 epochs with early stopping. Key features included global average pooling, two dense layers, dropout layers, and a softmax output layer for multi-class emotion classification.

### 7.5 VGG16 (Model 5)

Model 5, built upon the VGG16 architecture pre-trained on ImageNet, involved fine-tuning the base model with additional dense layers. The training process was performed with the Adam optimizer and categorical cross-entropy loss. Early stopping was implemented for model convergence over 40 epochs. The model architecture consisted of global average pooling, dense layers, and dropout layers, culminating in a softmax output layer for emotion classification.

In each training process, our models were assessed on the training and validation datasets. Additionally, early stopping played a pivotal role in preventing overfitting. The results, metrics, and model performance are explored in subsequent sections, enabling the selection of the optimal model for facial emotion recognition.

## 8 Model Evaluation

Let's start with an overview of the performance metrics for each of the tested models:

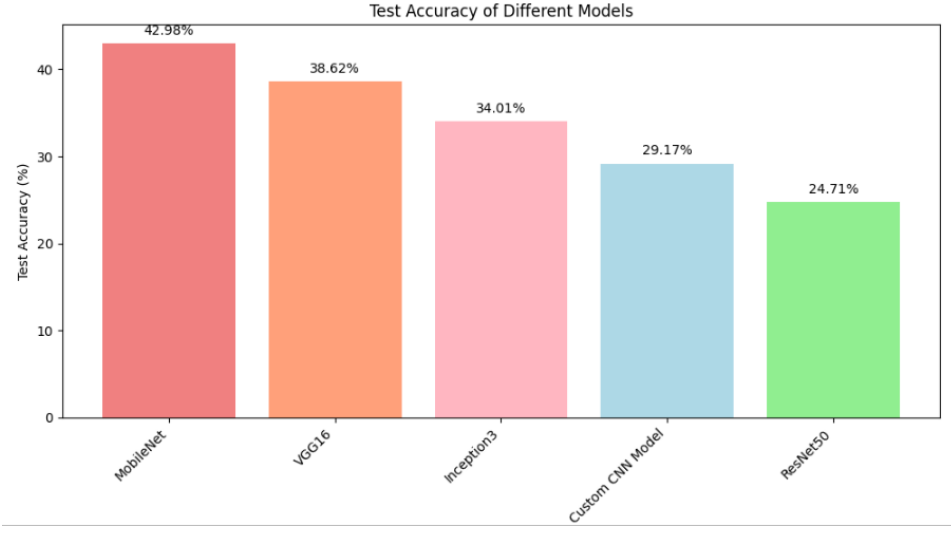


Figure 3: Test Accuracy of Different Models

Upon closer inspection, it becomes evident that the MobileNet model stands out as the top performer among all models. It achieved the highest test accuracy, F1 score, and demonstrated balanced precision and recall. This highlights the suitability of MobileNet for emotion classification tasks in image data.

Let's take a look at the training and validation scores for the MobileNet model.

We can observe that the training and validation loss gradually decrease over time, while training and testing accuracy increase. This indicates that our model is neither overfitting nor underfitting the data.

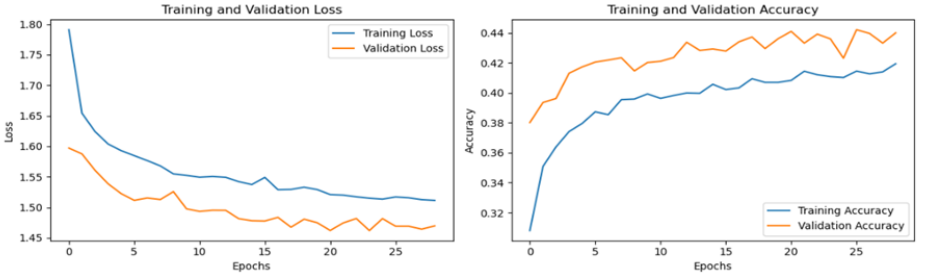


Figure 4: MobileNet : Training vs Validation

Let's take a closer look at the confusion matrix for the MobileNet model

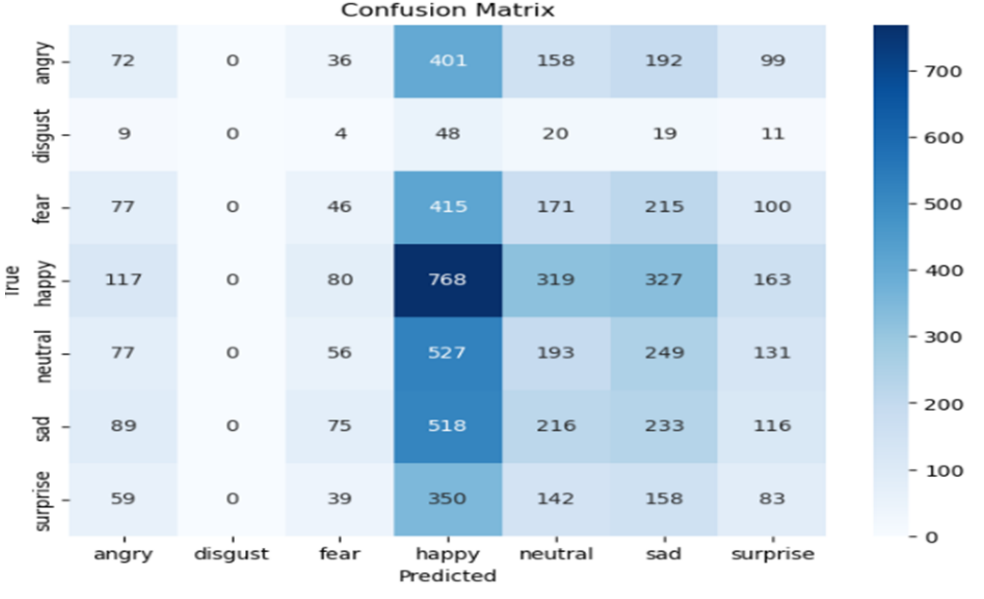


Figure 5: Confusion matrix of MobileNet

The confusion matrix further illustrates the MobileNet model's proficiency in classifying emotions. However, it's essential to acknowledge that the "disgust" class poses a significant challenge, with a precision and recall of 1.00 and 0.00, respectively. This suggests that the model struggles to correctly classify "disgust" images.

To gain a better understanding of the challenges posed by the "disgust" class, we conducted a more in-depth analysis. Upon examining the dataset, we discovered that "disgust" images are particularly challenging, even for human observers. Their features might be subtle and challenging to distinguish.

Considering the low performance in classifying "disgust" images, we recommend focus-

ing on improving the dataset quality and quantity for this specific class. Acquiring more diverse and representative "disgust" images and conducting data augmentation techniques can enhance the model's capability to correctly classify this emotion.

# 9 Challenges

Overcoming the challenges, we encountered in this project required us to address several crucial aspects. Long training times for deep learning models demanded patience, while limited computational resources posed some constraints. The process of choosing the right model for facial emotion recognition proved to be complex, and data imbalance in the dataset needed careful handling to ensure fairness.

# 10 Future Work

Moving forward, we aim to explore more advanced CNN architectures like EfficientNet, Vision Transformers (ViT), or DenseNet [20] to enhance our models' capabilities. Combining CNNs with other model types in hybrid configurations holds promise for improved performance. We plan to conduct hyperparameter tuning to optimize our models and increase accuracy. By applying data augmentation techniques and expanding our dataset, we seek to boost model generalization. Lastly, continuous evaluation and adjustment of our CNN architectures can lead to even better results.

In conclusion, recognizing these challenges and embracing future work can help us build more accurate and robust facial emotion recognition models, ensuring better results and wider practical applications.

# References

[1] Giannopoulos, P., Perikos, I. and Hatzilygeroudis, I. (2017). Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013. Advances in Hybridization of Intelligent Methods, pp.1–16. doi: https://doi.org/10.1007/978-3-319-66790-$4_1$.

[2] Liew, C.F. and Yairi, T. (2015). Facial Expression Recognition and Analysis: A Comparison Study of Feature Descriptors. IPSJ Transactions on Computer Vision and Applications, 7(0), pp.104–120. doi: https://doi.org/10.2197/ipsjtcva.7.104.

[3] Harms, M.B., Martin, A. and Wallace, G.L. (2010). Facial Emotion Recognition in Autism Spectrum Disorders: A Review of Behavioral and Neuroimaging Studies. Neuropsychology Review, [online] 20(3), pp.290–322. doi: https://doi.org/10.1007/s11065-010-9138-6.

[4] Zahara, L., Musa, P., Prasetyo Wibowo, E., Karim, I. and Bahri Musa, S. (2020). The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi. [online] IEEE Xplore. doi: https://doi.org/10.1109/ICIC50835.2020.9288560. [5] Durga, B.K. and Rajesh, V. (2022). A ResNet deep learning based facial recognition design for future multimedia applications. Computers and Electrical Engineering, 104, p.108384. doi: https://doi.org/10.1016/j.compeleceng.2022.108384.

[6] Moravčík, E. and Sebastián Basterrech (2021). Image-Based Facial Emotion Recognition Using Convolutional Neural Networks and Transfer Learning. Springer eBooks, pp.3–14. doi https://doi.org/10.1007/978-3-030-87178-$9_1$.

[7] Sujanaa, J., Palanivel, S. and Balasubramanian, M. (2021). Emotion recognition using support vector machine and one-dimensional convolutional neural network. Multimedia Tools and Applications. doi: https://doi.org/10.1007/s11042-021-11041-5.

[8] Wang, F., Lv, J., Ying, G., Chen, S. and Zhang, C. (2019). Facial expression recognition from image based on hybrid features understanding. Journal of Visual Communication and Image Representation, 59, pp.84–88. doi: https://doi.org/10.1016/j.jvcir.2018.11.010.

[9] k, hara (2016). Deep neural networks and regression models for object detection and pose estimation - ProQuest. [online] www.proquest.com. Available at: https://www.proquest.com/o
origsite=gscholarcbl=18750. [Accessed 29 Oct. 2023]

[10] Kusuma, G.P., Jonathan, J. and Lim, A.P. (2020). Emotion Recognition on FER-2013 Face Images Using Fine-Tuned VGG-16. Advances in Science, Technology and Engineering Systems Journal, [online] 5(6), pp.315–322. doi: https://doi.org/10.25046/aj050638.

[11] Dubey, A.K. and Jain, V. (2020). Automatic facial recognition using VGG16 based transfer learning model. Journal of Information and Optimization Sciences, 41(7), pp.1589–1596. doi: https://doi.org/10.1080/02522667.2020.1809126.

[12] Chowdary, M.K., Nguyen, T.N. and Hemanth, D.J. (2021). Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications. doi: https://doi.org/10.1007/s00521-021-06012-8.

[13] Amornvit Vatcharaphrueksadee, Rattikarn Viboonpanich, Puttakul Sakul-ang and Maleerat Maliyaem (2020). VGG-16 and Optimized CNN for Emotion Classification. Information Technology Journal, [online] 16(2), pp.10–15. Available at: https://ph01.tci-thaijo.org/index.php/IT*journal*/*article*/*view*/243769[*Accessed*29*Oct*.2023].

[14] Akhand, M.A.H., Roy, S., Siddique, N., Kamal, M.A.S. and Shimamura, T. (2021). Facial Emotion Recognition Using Transfer Learning in the Deep CNN. Electronics, 10(9), p.1036. doi: https://doi.org/10.3390/electronics10091036.

[15] Xia, X.-L., Xu, C. and Nan, B. (2017). Facial Expression Recognition Based on TensorFlow Platform. ITM Web of Conferences, 12, p.01005. doi: https://doi.org/10.1051/itmconf/2

[16] Islam Nur, A., Iman H, K. and Pandu, W. (2022). Transfer Learning Technique with EfficientNet for Facial Expression Recognition System. Revue d'Intelligence Artificielle., pp.p543-552.

[17] Li, B. and Lima, D. (2021). Facial expression recognition via ResNet-50. International Journal of Cognitive Computing in Engineering, 2, pp.57–64. doi: https://doi.org/10.1016/j

[18] Vepuri, K. (2021). Improving Facial Emotion Recognition with Image processing and Deep Learning. Master's Projects. [online] doi: https://doi.org/10.31979/etd.3wrz-53ee.

[19] Carrier, P.-L. and Courvill, A. (2013). FER-2013. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/msambare/fer2013.

[20] Nan, Y., Ju, J., Hua, Q., Zhang, H. and Wang, B. (2022). A-MobileNet: An approach of facial expression recognition. Alexandria Engineering Journal, 61(6), pp.4435–4444. doi: https://doi.org/10.1016/j.aej.2021.09.066.

[21] Ketkar, N. and Moolayil, J. (2021). Convolutional Neural Networks. Deep Learning with Python, pp.197–242. doi: https://doi.org/10.1007/978-1-4842-5364-$9_6$.

[22] Meena, G., Mohbey, K.K., Kumar, S., Chawda, R.K. and Gaikwad, S.V. (2023). Image-Based Sentiment Analysis Using InceptionV3 Transfer Learning Approach. SN Computer Science, 4(3). doi: https://doi.org/10.1007/s42979-023-01695-3.

[23] Mandal, B., Okeukwu, A. and Theis, Y. (2021). Masked Face Recognition using ResNet-50. arXiv:2104.08997 [cs]. [online] Available at: https://arxiv.org/abs/2104.08997.

[24] Theckedath, D. and Sedamkar, R.R. (2020). Detecting Affect States Using VGG16, ResNet50 and SE-ResNet50 Networks. SN Computer Science, 1(2). doi: https://doi.org/10.1007/020-0114-9.