A Comparative Study on the Effects of Pooling on FER CNN Models

Muskan Agrawal
Department of Computer Science
BIT Raipur
India

Padmavati Shrivastasva
Department of Computer Science
BIT Raipur
India

Rahul R. Pillai
Department of Computer Science
BIT Raipur
India

Chirag Budhwani
Department of Computer Science
BIT Raipur
India

Shivam Khare
Department of Computer Science
BIT Raipur
India

ABSTRACT

Emotion recognition has attracted much attention in Artificial Intelligence in order to make machines understand emotional sentiments, with many industries trying hard to incorporate emotion recognition technologies into their products. The easiest way to detect a person’s emotion is recognizing their facial expressions. In this work, the researchers tend to use FER as a problem and use the Deep Convolutional Neural Network (DCNN), which extracts the features automatically and therefore surpasses and outperforms the limitations of traditional machine learning. This work provides a comparative study of various existing pre-trained model architectures. MobileNetV2, MobileNetV3_Small, NASNetMobile, ResNet50, ResNet50V2, ResNet152V2, DenseNet169 and DenseNet201 with modification in their pooling layer to achieve high accuracy and have the potential for implementation in embedded systems. In this project, various deep learning pre-trained models were trained, tested, and compared on a modified subset of the FER 2013 Dataset for Face Emotion Recognition under all the conditions of pooling, i.e., None, Min, Avg, and Max. FER 2013 being one of the most challenging dataset, and due to limited run-time cost available, the MobileNetV2 model gave the highest testing accuracy of 83.64% with a training accuracy of 97.87% on average pooling.

The models were compared on the following evaluation metrics: Accuracy, Loss, Precision, Recall and F1-score. For a practical approach, they integrate the model into a mobile application so that models can be run on devices in real time.

General Terms

Artificial Intelligence, Computer Vision, Image Recognition

Keywords

Transfer Learning, Pooling, Face Emotion Recognition

1. INTRODUCTION

Computer vision, a science of artificial intelligence, works on image processing and visual-related analysis which enables computers to work in a comparable manner to human vision. It entails the development of ways to mimic the capabilities of human eyesight. The objective of computer vision is to outperform biological vision at extracting valuable information from visual data. These computer vision elements were chosen by the researchers to recognise a human's facial expression and hence his or her mood.

Facial Emotion Recognition is a computer vision application intended to detect varied facial expressions in human beings in order to determine an individual's emotion as facial expressions are the primary source of a person's mood. Because of its capacity to replicate human coding abilities, an AI-based facial expression recognition system is crucial. Nonverbal communication signals such as facial expressions play a vital part in interpersonal relationships. These signals support speech by assisting the listener in deciphering the meaning of uttered words. Face expression recognition is able to give unfiltered, impartial emotional reactions as data since it takes and analyses information from an image or video source.

In the following research, the goal is not to bolster the efficiency and better understand the emotion recognition models but to make the best use of them in real-life scenarios. Here, a modified subset of the FER2013 dataset is being used. The dataset is divided into 2 parts: Training set (7500 images) and Testing set (5000 images) classified into 5 different emotions. The concept of transfer learning was applied to overcome the gaps caused by the unavailability of huge amounts of data and computational resources. Following are the pre-trained models used in this experiment:

- MobileNet_V2
- MobileNetV3_Small
- ResNet50
- ResNet50V2
- ResNet152V2
- NASNetMobile
- DenseNet169
- DenseNet201

In this paper, the researchers start by giving an introduction to the project, followed by Section 2: Related Works, in which they state their findings from various research and reference papers. Section 3 describes the prerequisite. Section 4 defines the methodology of the project. Section 5 showcases the observations and results of the compared models. Section 6 presents the conclusion reached as a result of them. Section 7 highlights the future scope of the project.

2. RELATED WORKS

In [3] Amil Khanzada et al. have attempted to enhance the results of pre-trained models and to implement that model in the real world via the mobile web app. The models were trained using transfer learning to address high variance batch normalization, dropout layers, softmax layer, and max pooling performed.They got an accuracy of 75.8% on the FER2013 dataset after using an enhanced dataset.

In [4], Hung et al. sought to enhance the FaceLiveNet network accuracy in fundamental expression recognition and developed the Dense FaceLiveNet framework employing JAFFE, KDFC, and FER2013 datasets. The model was built using 5-fold cross-validation, with Swish replacing the ReLU. As a result, the authors were able to effectively develop two-phase transfer learning to establish the emotion detection model and bridge the difficulty of limited data for emotion analysis.

Martina Rescigno et al. discovered in [5] that generic transferred knowledge combined with a minimal quantity of personal data is enough to achieve great detection performances and improvement for both a universal model and a personalized model. The comparison was done using three models: a self-trained model on AffectNet; a model trained using transfer learning from predefined model AlexNet and a model trained with transfer learning and finely tuned with a personal database.

Lida Hu and Qi Ge introduced an approach in [6] that integrates a transfer learning method with a joint supervision technique with island loss, which is important for face tasks. They created MobileNetV2, that has achieved state-of-the-art accuracy on the JAFFE and CK+ datasets (i.e. 97.98 percent and 95.238 percent, respectively) and a 73 percent accuracy on the FER2013 dataset.

In [7] Arjun Singh et al. constructed a model for the study of sensory learning. The model scored 67.20% in the Kaggle dataset and 78.32% in the KDEF dataset with the transfer practice.

In [8], Nadir Kamel Benamara et al. proposed an emotion identification system for friendly robots based on a YOLO-based face detection system and an ensemble CNN. Different models A, B, C, and D were developed using the FCNN approach, batch normalization, and dropout layer. These models were ensemble in all possible manners and the ensemble model ABC and ABCD give a high performance of 72.47% in the FER2013 dataset.

Li and Dias Lima [9] created a feature extraction approach for face emotion identification utilizing the deep residual network ResNet-50, which incorporates convolutional neural networks. To increase the model's convergence capabilities, batch normalization and ReLU were applied. The model has an impressive accuracy of 95.39±1.41%.

To overcome the constraints of established CNN models, Abhinav Agrawal and Namita Mittal in [10] utilized the FER-2013 dataset to develop new CNN models by evaluating alternative kernel sizes and filter counts. These designs are basic and distinct in terms of hyper-parameter selection across network layers, and they may be used to standardize the base model for the widely contested FER-2013 dataset. They employed a component layer as a building element of the network, which is a composition of kernel size and filter count. According to the findings of their research, kernel size and filter count have a substantial influence on the network's accuracy.

In [11] Luan Pham et al. suggested a unique Masking Idea to improve performance. It modifies feature maps using a segmentation network, which allows the network to focus on important data and make right judgments. They created a Residual Masking Network for FER. Their Residual Masking Network ensembles with 6 CNNs achieved the highest score of 76.82 percent in the ensemble mode, while their single Residual Masking Network received the highest mark of 74.14 percent on a new dataset called VEMO.

In [12], the goal was to optimize the accuracy-latency trade-off of computer vision architecture for mobile devices. Andrew Howard et al. developed highly efficient MobileNetV3 Large and MobileNetV3 Small models in order to offer the next generation accuracy of neural network models for computer vision. The models are built by utilizing platform-aware NAS and NetAdapt for network search and adding network optimizations. MobileNetV3 Large and Small performance trade-offs as a function of multiplier and resolution are 73.8 % and 64.9 %, respectively. MobileNetV3-Small beats MobileNetV3-Large by roughly 3% when multipliers are adjusted to equal performance.

In [13]Barret Zoph et al. employ the NAS framework, a reinforcement learning search approach to enhance the architecture configurations. They used a transferability approach to transfer the learned block from sub-dataset to the complete dataset. They achieved this transferability by creating a search space (called "the NASNet search space") in which the complexity of the architecture is independent of network depth and image size. On ImageNet, the model achieved accuracy of 82.7 % top-1 and 96.2 % top-5 on the CIFAR-10 dataset.

In [14], Gao Huang Et Al. established a novel design, DenseNet, with nearly half the connection length of any existing convolutional network. DenseNets offer numerous appealing advantages: tThey effectively address the issue of vanishing gradients, boost feature propagation, maximize feature reuse, and drastically cut down on the required number of parameters.

In [15] Kaiming He et al. overcame the deterioration problem by developing a deep residual learning approach. They presented a theoretical method stating that optimizing the residual mapping is easier than optimizing the original, unreferenced mapping. As a result of their research, they find that highly in-depth residual nets are simple to optimize, whereas "flat" nets display increased error as the depth grows.

In [16] Joe Wang et al. had performed transfer learning on three known CNNs: InceptionV3, MobileNet, and NASNet. They used 14000 images for their dataset drawn from the extended CK+(Cohn-Kanade) dataset and Japanese Female Facial Expressions(JAFFE). Data augmentation techniques were used. These included: random flipping, random cropping, random scaling, and random brightness adjustments on images. They reported the training accuracy of 51.59% of NASNet, 60.91% of MobileNet, and 74.04% of InceptionV3.

In [17] Atharav Godhkar et al. had performed a comparative study on numerous pre-trained models to recognise emotion through facial expression. The models for the study were trained using transfer learning over the CK+ dataset. Considering the inadequate size of the dataset, data augmentation: Rescale, Shear, Zoom and Horizontal Flip were performed. They obtained the validation accuracy higher than 90% for ResNet50V2, ResNet101V2, ResNet152V2, and MobileNet. MobileNet was selected as the most efficient model in their research due to its comparatively smaller size and higher accuracy.

In [18] George-Cosmin Porușniuc et al. investigated three convolutional networks: a basic serial design, a network inspired by the Xception structure, and the ResNet50 model. Two more strategies were applied to improve the performances. To begin, several ensembles of models were constructed by repeatedly training the networks on varied versions of the training subset produced using a "bagging" approach. Second, they had employed transfer learning to fine-tune the settings. They obtain ResNet50 trained with transfer learning to be the most accurate model with an accuracy of 71.5% on the FER2013 dataset. They even concluded that the ensemble of models increases the accuracy of the model.

To enhance the results of FER and improvise the generalization of the training model, Lei Yang et al. in [19] modified the traditional Inception-ResNet-v1 by replacing the existing activation function(softmax) with Support Vector Machine (SVM). The improvised model gave an accuracy of 99.6% and 68.1% on the CK+ dataset and FER2013 dataset respectively.

In [20] M. A. H. Akhand et al. conducted a comparison investigation utilizing eight pre-trained models: VGG-16, VGG-19, ResNet-18, ResNet-34, ResNet-50, ResNet-152, Inception-v3, and DenseNet-161. To make it compatible with FER, the pre-trained DCNN models for image classification were adopted and trained using transfer learning by substituting its top layers with the dense layer(s). The proposed method for DenseNet161 has outshined the other pre-trained models used in the study. The model achieved an accuracy of 99.52% in 10-Fold CV on JAFFE and regarding the KDEF dataset, achieved an accuracy of 98.78%.

3. PRE-REQUISITES

3.1 Convolutional Neural Networks

In artificial intelligence, computer vision and image processing are based on convolutional neural networks(CNNs). It has received a lot of traction among academics for constructing a FER system because of its outstanding performance in learning characteristics. Convolutional neural networks, in general, are composed of numerous layers of tiny neurons that convert the input image into receptive fields. It has an input layer, an output layer, and a number of hidden layers. Convolutional layers, fully connected layers, pooling layers, and normalizing layers(ReLU) are all parts of a CNN's internal architecture. For increasingly sophisticated models, more layers might be implemented.

The convolutional layer uses a series of filters to convolve over the input image and generate several kinds of feature maps. Local connection, weight sharing, and shift-invariance are the three fundamental advantages of the convolution process.

In order to reduce the network's computational cost, a pooling layer is used after a convolutional layer to shrink the feature maps' spatial dimensions [21]. The pooling layer, also known as the down-sampling layer, tries to lower the size of a convolutional layer's matrix. On the one hand, pooling decreases the number of features and parameters, making convolutional network computation more straightforward. It also helps to prevent overfitting to some extent, making optimization easier. On the other hand, some properties, such as rotation, translation, and contraction, are preserved. There are three types of pooling: maximum, average and minimum.

3.1.1 Max Pooling:

Max pooling is a mathematical procedure that extracts the greatest value from a specified region of an image with a definite size.

3.1.2 Average Pooling:

Average pooling is a mathematical procedure that calculates the average value of a region of an image of a specific size [22].

3.1.3 Min Pooling:

Min pooling is a mathematical process that takes the smallest value of a section of an image of a particular size.

A CNN's normalization layer entails setting all negative values in the filtered image to 0 in the normalization layer. It improves the model's nonlinear features.

3.2 Deep Convolutional Neural Networks

A DCNN model is made up of many hidden layers and uses high-dimensional images, making input and training difficult. In the convolutional layers, each DCNN model has its own important layouts and connections. Because the DCNN model has a huge set of parameters to modify, training a new model is a difficult operation. Because it is such a big and complex network, training it necessitates a large dataset. Training using a big dataset may get the desired result, but it would be time expensive, and a lack of data could lead to over-fitting. It's challenging to gather enough information for each activity. Because it isn't always easily available in some circumstances. But, researchers have found that transfer learning can be used to overcome this issue.

3.3 Transfer Learning

The technique of transferring information from past training to be utilized in new research in order to shorten training time is known as transfer learning. This is quite different from standard machine learning, which requires a considerable computing period and requires learning input data from the beginning. Transfer learning is a method for reusing a previously learned CNN model, often known as a pre-trained model. The learnt features from the pre-trained models are then transferred to the proposed dataset for training.

3.4 Pre-trained Models

3.4.1 MobileNetV2

The Inverted Residual Block, which uses ResNet's short-cut design and a mix of a depthwise convolution and a 1 x1 pointwise convolution, is at the base of MobileNet V2. This FER job is able to maintain a positive trade-off between speed and precision due to its compact size, fast running speed, and outstanding accuracy[6]. MobileNetV2[11] adds a resource-efficient block with inverted residuals and linear bottlenecks to MobileNetV1, having one complete 32 filters' convolutional layer and 19 residual bottleneck layers.

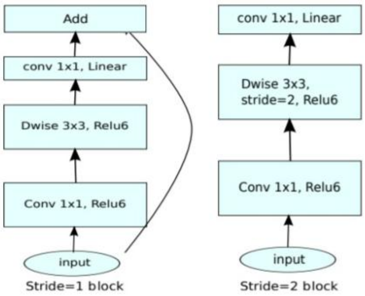

Figure 1: Architecture of MobileNetV2

3.4.2 MobileNetV3:

The architecture of MobileNetV3 is more advanced. Squeeze-and-excitation blocks were introduced in MobileNetV2's network, resulting in a more robust architecture.

The MobileNetV3-Large and MobileNetV3-Small are the two available models. These models are intended for both high and low resource utilization scenarios. Platform-aware NAS, NetAdapt, and network search are used to build the models together with network upgrades. The sigmoid is used in both squeezing and excitation, as well as the swish nonlinearity, but it is inefficient to compute and difficult to maintain accuracy in fixed-point arithmetic, therefore they utilize the hard sigmoid [12].

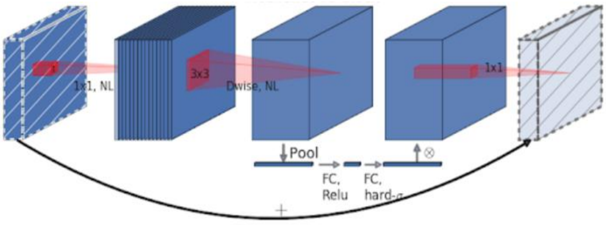

Figure 2: Architecture of MobileNetV3

3.4.3 NASNetMobile

The acronym NASNet stands for Neural Architecture Search Network. Normal cell and Reduction cell are the two basic functions used.

● Normal Cell: Convolutional cells that produce a feature map of the same dimension are known as normal cells.

● Reduction Cell: Convolutional cells that yield a feature map with the height and breadth of the feature map decreased by a factor of two.

Although both the cells might have the same architecture, it's found that learning two different architectures was more advantageous. When the spatial activation size is lowered, a common heuristic is to double the filter count in the output to preserve a roughly constant hidden state dimension [13].

To obtain the higher map, NASNet first performs its operations to the small dataset and then transfers its block to the large dataset. The normal cells set the feature map size, while the reduction cell delivers the reduced feature map in terms of height and breadth by a factor of two in the original NASNet Architecture, where the number of cells is not predefined and only normal and reduction cells are employed. In NASNet, a Recurrent Neural Network (RNN)-based control architecture is used to forecast the complete structure of the network based on the two initial hidden states.

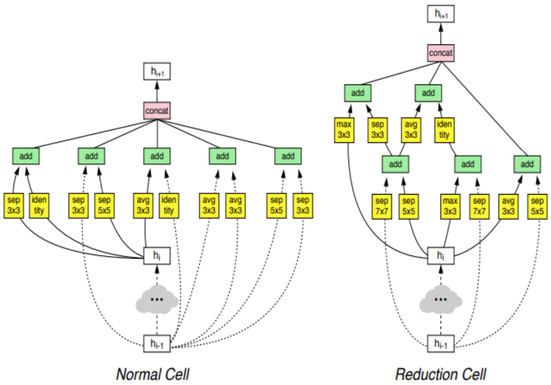

Figure 3: Architecture of NASNet

3.4.4 ResNet50

ResNet50, an upgrade from ResNet34, is constructed by substituting each 2-layer block with a 3-layer bottleneck block, hence increasing the number of layers in the network from 34 to 50. To achieve classification tasks, ResNet50 conducts convolution on the input layer at the start, which is followed by four residual blocks, and lastly, a fully connected operation. Figure 4 shows the ResNet-50 network topology, which contains 50 Conv2D operations [9].

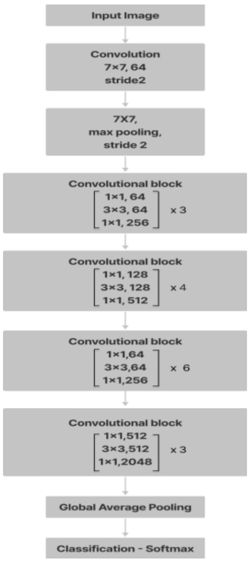

Figure 4: Architecture of ResNet50

3.4.5 ResNet152V2

It consists of 152 layers that use batch normalization before each layer [14].

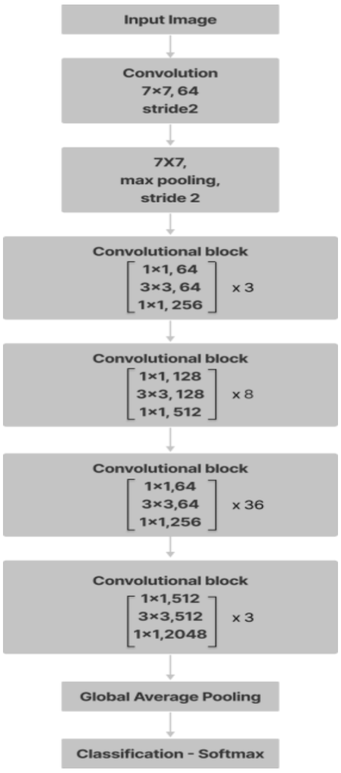

Figure 5: Architecture of ResNet152

3.4.6 DenseNet169

DenseNet169 features 169 depth layers, a small number of parameters compared to other models, and handles the vanishing gradient problem better [17].

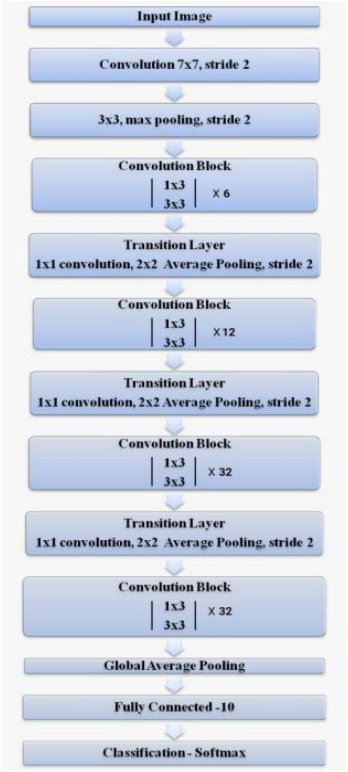

Figure 6: Architecture of DenseNet169

3.4.7 DenseNet201

DenseNet201 makes use of the compacted network to produce models that are simple to train and very efficient in terms of network parameters by reusing features across layers. This improves performance because each following layer can draw on all the feature maps from all the preceding layers, increasing the input diversity and enhancing generalization [17].

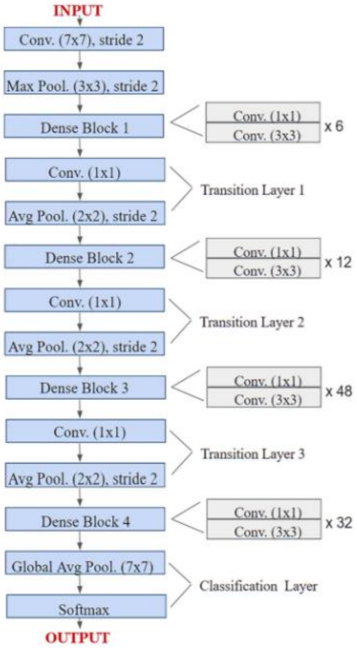

Figure 7: Architecture of DenseNet201

4. METHODOLOGY

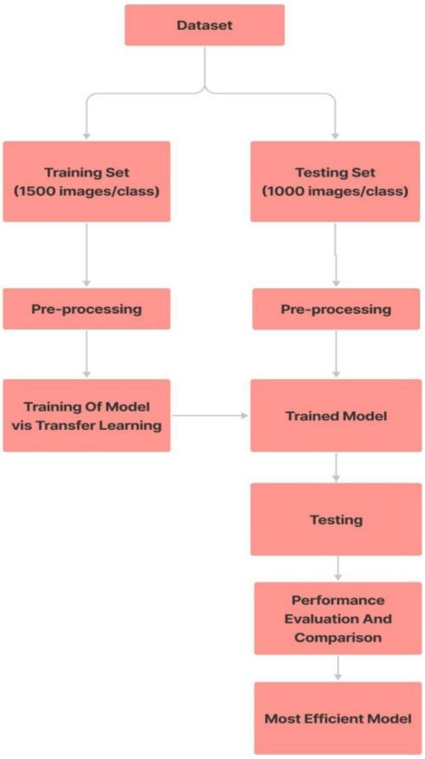

Figure 8: Flowchart of Methodology

4.1 Dataset Description

The FER 2013 dataset was presented in ICML 2013 workshop on “Challenges in representation learning”. The dataset has two parts: Training set (28709 images) and Testing set(3589 images) which consists of grayscale face images of 48x48 per resolution classified into 7 different emotions: angry, sad, neutral, happy, surprise, fear, and disgust[23].

In an attempt to improve the result of the model and surpass the limitations of the FER2013 dataset, a balanced dataset is required and therefore a modified subset of FER2013 is used. The sample dataset consists of 5 basic classes: Angry, Anxious, Happy, Sad, and Neutral. Since Disgust had a comparatively small number of images, we dropped that class as it was reducing the accuracy of the models. Also, we merged fear and surprise to create a new class Anxious. The modified subset is then divided into two parts: The training sample with 7500 images (1500 images per emotion category) and the Testing sample with 5000 images (1000 images per emotion category).

4.2 Pre-processing

Preprocessing aims to resize and normalize the images of the dataset as required by each models’ architecture. Images are resized from 48 × 48 px to 224 x 224px to match the input layer of pre-trained models followed by normalization of each pixel by dividing with 255.0 uniformly for every image so as to define a fixed [0, 1] normalization range.

4.3 Training the model

For all the pre-trained models used in the experiment, the researchers have configured the type of pooling for each of the existing pooling layers as per model architecture. Then the output layer for every model is changed from 1000 classifications to 5 classifications with a softmax activation function. Then, each network is trained for 50 iterations (epochs) over the training set with sparse_categorical_crossentropy loss and Adam optimizer.

4.4 Testing the model

Models are tested over the testing dataset which gives us a testing accuracy and overall loss.

4.5 Performance Evaluation

Each tested model is then evaluated on the following evaluation metrics: Accuracy, Loss, Precision, Recall and F1-Score. The comparison of models on all these metrics can be found in the observation table 5.5.

5. RESULT

After training the models with the configurations mentioned in section 4, they are compared on the basis of Type of Pooling, Model Size(in mb), their Accuracy, and their Loss percentages observed on training and testing data and noted in the Observation Tables given below:

Table 1: Observation Table for None Pooling

Model Name	Model Size	Training Accuracy	Training Loss	Testing Accuracy	Testing Loss
MobileNet_V2	28.4	97.87	06.20	44.16	5.17
MobileNetV3_Small	31.7	98.13	05.43	19.94	4.46
NasNetMobile	54.0	97.87	06.23	32.14	3.15
ResNet50	273.6	98.75	03.55	55.00	2.47
ResNet50V2	273.3	98.40	03.70	56.00	2.46
ResNet152V2	672.2	97.72	06.50	48.44	2.65
DenseNet169	148.8	98.00	06.12	58.62	2.06
DenseNet201	214.2	97.75	01.87	58.32	2.02

Table 2: Observation Table for Max Pooling

Model Name	Model Size	Training Accuracy	Training Loss	Testing Accuracy	Testing Loss
MobileNet_V2	28.4	97.16	8.17	53.00	2.92
MobileNetV3_Small	31.7	97.33	8.37	20.66	3.69
NasNetMobile	54.0	97.02	8.95	28.60	10.19
ResNet50	273.6	98.92	3.41	54.66	2.34
ResNet50V2	273.3	98.28	4.69	55.48	2.66
ResNet152V2	672.2	98.64	3.91	54.83	2.79
DenseNet169	148.8	98.28	5.63	58.80	2.33
DenseNet201	214.2	98.27	2.37	57.86	2.52

Table 3: Observation Table for Average Pooling

Model Name	Model Size	Training Accuracy	Training Loss	Testing Accuracy	Testing Loss
MobileNet_V2	28.4	97.87	02.88	83.64	7.19
MobileNetV3_Small	31.7	96.46	14.64	22.60	2.87
NasNetMobile	54.0	98.45	04.68	22.12	3.08
ResNet50	273.6	98.28	03.53	74.64	9.80
ResNet50V2	273.3	98.95	03.14	55.56	2.56
ResNet152V2	672.2	97.41	09.02	48.06	2.98
DenseNet169	148.8	97.53	06.99	58.84	2.03
DenseNet201	214.2	98.21	04.07	59.28	2.31

Table 4: Observation Table for Min Pooling

Model Name	Model Size	Training Accuracy	Training Loss	Testing Accuracy	Testing Loss
MobileNet_V2	28.4	97.65	5.46	52.86	2.95
MobileNetV3_Small	31.7	97.95	6.13	19.08	5.63
NasNetMobile	54.0	98.37	4.37	34.58	5.51
ResNet50	273.6	98.77	3.50	54.32	2.44
ResNet50V2	273.3	98.64	3.46	53.64	2.56
ResNet152V2	672.2	97.48	7.79	51.27	2.72
DenseNet169	148.8	97.88	6.40	60.34	2.35
DenseNet201	214.2	98.27	2.88	59.28	2.16

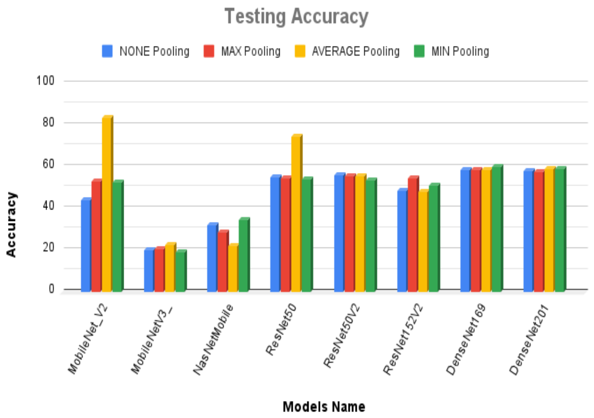

Figure 9: Graph on testing accuracies

From the observation table and graph, it can be noted that MobileNetV2 has the maximum testing accuracy of 83.64% with a training accuracy of 97.87% on average pooling.

Whereas ResNet50 gave a training accuracy of 98.92% on max pooling which was the highest but the testing accuracy was 54.66%.

Even though some research studies surpass this accuracy, those procedures either do not implement a real-time application or use much more costly or upgraded run-time resources than what was available in this research.

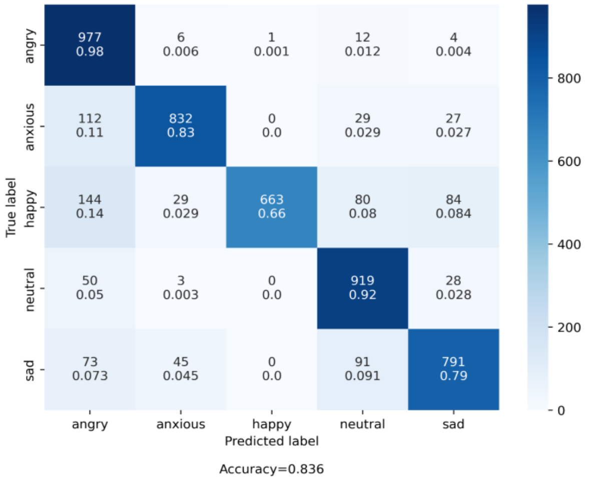

Figure 10: Confusion matrix for the best model

Table 5: Evaluation Metrics

	Precision	Recall	F1-Score	Support
Angry	0.720501	0.977000	0.829372	1000
Anxious	0.909290	0.832000	0.868930	1000
Happy	0.998494	0.663000	0.796865	1000
Neutral	0.812555	0.919000	0.862506	1000
Sad	0.846895	0.791000	0.817994	1000
Accuracy	0.836400	0.836400	0.836400	0.836400

6. CONCLUSION

In this project, numerous deep learning pre-trained models were trained, tested, and as compared on a modified subset of the FER2013 Dataset for Face Emotion Recognition under all the conditions of pooling i.e. None, Min, Avg, and Max.

7. FUTURE SCOPE

To further increase the efficiency of the models to be used in real-life cases, a diversified dataset with various racial origins can be used together. This work can be expanded by implementing the emotion recognition model in a Mental Health Tracker App to detect the emotions using facial expression along with psychological aspects to improve user experience. Also to enhance the emotion recognition, a multi-modality approach such as visual, textual, audio etc can be used.

8. REFERENCES

[1] Nandhini Abirami, R., Durai Raj Vincent, P. M., Srinivasan, K., Tariq, U., & Chang, C. Y. (2021). Deep CNN and Deep GAN in Computational Visual Perception-Driven Image Analysis. Complexity, 2021.

[2] De Silva, L. C., Miyasato, T., & Nakatsu, R. (1997, September). Facial emotion recognition using multi-modal information. In Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat. (Vol. 1, pp. 397-401). IEEE.

[3] kay, F. T. (2020). Facial expression recognition with deep learning. arXiv preprint arXiv:2004.11823.

[4] Hung, J. C., Lin, K. C., & Lai, N. X. (2019). Recognizing learning emotion based on convolutional neural networks and transfer learning. Applied Soft Computing, 84, 105724

[5] Rescigno, M., Spezialetti, M., & Rossi, S. (2020). Personalized models for facial emotion recognition through transfer learning. Multimedia Tools and Applications, 79(47), 35811-35828.

[6] Hu, L., & Ge, Q. (2020, June). Automatic facial expression recognition based on MobileNetV2 in Real-time. In Journal of Physics: Conference Series (Vol. 1549, No. 2, p. 022136). IOP Publishing.

[7] Singh, A., Srivastav, A. P., Choudhary, P., & Raj, S. (2021, April). Facial emotion recognition using convolutional neural networks. In 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM) (pp. 486-490). IEEE.

[8] Benamara, N. K., Val-Calvo, M., Álvarez-Sánchez, J. R., Díaz-Morcillo, A., Vicente, J. M. F., Fernández-Jover, E., & Stambouli, T. B. (2019, June). Real-time emotional recognition for sociable robotics based on deep neural networks ensemble. In International Work-Conference on the Interplay Between Natural and Artificial Computation (pp. 171-180). Springer, Cham.

[9] Li, B., & Lima, D. (2021). Facial expression recognition via ResNet-50. International Journal of Cognitive Computing in Engineering, 2, 57-64.

[10] Agrawal, A., & Mittal, N. (2020). Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. The Visual Computer, 36(2), 405-412.

[11] Pham, L., Vu, T. H., & Tran, T. A. (2021, January). Facial Expression Recognition Using Residual Masking Network. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 4513-4519). IEEE.

[12] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1314-1324).

[13] Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8697-8710).

[14] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[15] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[16] Mbuthia, J. W. M. FaceNet: Facial Expression Recognition based on Deep Convolutional Neural Networks.

[17] Gondkar, A., Gandhi, R., & Jadhav, N. (2021, October). Facial Emotion Recognition using Transfer Learning: A Comparative Study. In 2021 2nd Global Conference for Advancement in Technology (GCAT) (pp. 1-6). IEEE.

[18] Porușniuc, G. C., Leon, F., Timofte, R., & Miron, C. (2019, November). Convolutional neural networks architectures for facial expression recognition. In 2019 E-Health and Bioengineering Conference (EHB) (pp. 1-6). IEEE.

[19] Yang, L., Zhang, H., Li, D., Xiao, F., & Yang, S. (2021, September). Facial Expression Recognition Based on Transfer Learning and SVM. In Journal of Physics: Conference Series (Vol. 2025, No. 1, p. 012015). IOP Publishing.

[20] Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., & Shimamura, T. (2021). Facial Emotion Recognition Using Transfer Learning in the Deep CNN. Electronics, 10(9), 1036.

[21] BELAL, F. (2020). BENCHMARKING OF CONVOLUTIONAL NEURAL NETWORKS FOR FACIAL EXPRESSIONS RECOGNITION. Journal of Theoretical and Applied Information Technology, 98(18).

[22] Sun, J., Slang, S., Elboth, T., Greiner, T. L., McDonald, S., & Gelius, L. J. (2020). Attenuation of marine seismic interference noise employing a customized U- Net. Geophysical Prospecting, 68(3), 845-871.

[23] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013, November). Challenges in representation learning: A report on three machine learning contests. In International conference on neural information processing (pp. 117-124). Springer, Berlin, Heidelber.