TriceptionNet: An Ensemble Learning Based Approach For Facial Emotion Recognition

Abhishek M   
AIT – CSE  *Chandigarh University*  
Mohali , Punjab  *21BCS6748@cuchd.in*  
Shashank Raj  
AIT – CSE  *Chandigarh University*  
Mohali , Punjab *21BCS6016@cuchd.in*  
Aarti   
AIT -CSE  *Chandigarh University*  
Mohali, Punjab *aarti.e15380@cumail.in*

*Abstract*— *In this study, we present TriceptionNet, a novel hybrid learning model for facial expression recognition. Our method leverages the distinct strengths of EfficientNetB0, InceptionResNetV2, and ResNet50, combining the best of our knowledge to increase demand accuracy. Our training and evaluation are based on the FER2013 dataset, which comprises a diverse range of image records. According to our findings, EfficientNetB0 has a training accuracy of 65.24%, whereas InceptionResNetV2 and ResNet50 have accuracies of 61.69% and 63.04%, respectively. The combined model's overall accuracy of 67% demonstrates how various theories perform differently and highlights issues with face recognition. This study offers insight for the future development of cognitive theory and shows how group learning can enhance classification performance.*

Keywords - EmotionNetB0, InceptionResNetV2, ResNet50 , FER 2013, Ensemble Learning

# **Introduction**

# Understanding the human emotions using facial expressions is an area of research with many applications in fields like mental health assessment, human – computer interaction , security systems etc. As use of technology continues to grow day by day , the ability to accurately interpret user emotions through facial expressions has become crucial. However this is not an easy task , as expressions can vary from individual to individual.

Previously , traditional methods for emotion recognition used manually crafter features and basic classification algorithms. While these techniques did work up to a certain extent, they struggled when it came to real world scenarios. The emergence of deep learning models like Convolutional Neural Networks have completely changed this field. These model work exceptionally well at automatically learning features from input images, allowing for more robust feature extraction and learning.

In this paper, we introduce our model for FER , the TriceptionNet, a new ensemble model that combine the strength of 3 CNN architectures: EfficientNetB0, InceptionResNetV2 and ResNet50. By ensembling or combining these model , we aim to achieve improved accuracy and reliability when it comes to emotion recognition. We trained and tested our model using the FER 2013 dataset, which have over 30000 images of facial expressions of 7 different emotions.

From the model’s performance it was identified that ensemble learning was able to significantly improve the performance by utilizing the individual strengths of the base learners. This work was not only able to help advance the field of facial emotional recognition, but was also able to pave the path for further developments in the field of classification.

# **Related works**

During recent times, due to the development of deep learning techniques, Facial Emotional Recognition (FER) has been an area of great interest. Since traditional methods often relied on simple classifiers and manually created features, generalizing it across diverse datasets and complex real-world scenarios turned out to be a struggle. But due to recent advancements in the field of deep learning, models like Convolutional Neural Networks (CNN) have helped in improving the accuracy and robustness of FER systems by a significant amount.

The study done by Kahou et al. (2017) [1], introduced the utilization of CNN architecture trained on the FER2013 dataset, for working on FER. This turned out to be one of the foundational works done in the domain. It unlocked the potential in capturing different facial features associated with different emotions. Taking this finding into consideration, various researches had been done using a wide variety of approaches to enhance the performance of FER models. Zhang et al. (2018) [2] an ensembled that combined the CNN models to improve the accuracy. Similarly, Li et al. (2019) [3] built a model to focus on relevant facial features for emotional detection by integrating attention mechanisms into it.

Gao et al. (2020) [4] introduced a model to capture temporal dynamics in facial expressions. In it they proposed a hybrid model combining convolutional Neural Networks and Recurrent Neural Networks. This method worked well in capturing emotions from video sequences, displaying the effectiveness of using temporal dependencies in emotion recognition tasks .A variety of data augmentation techniques have also been tested to find methods that work well for model generalization. Guo et al. (2021) [5] tested different data augmentation techniques to find best one for emotion recognition tasks. This study gave insights into which augmentation techniques where effective in avoiding overfitting and improving model performance.

Transfer learning has also been used as a way to improve performance in FER tasks. Mollahosseini et al. (2017) [6] used various pretrained models like VGG-16 and Resnet 50 which also showed considerable improvements in performance.

Approaches involving ensemble learning have also been used to improve performance of models. Kim et al. (2022) [7] came up with an ensemble model which showed improved performance when ensembling various CNN architectures.

In this paper, the system proposed which uses an MLP as meta classifier to ensemble an EfficientNetB0 , Resnet50 and InceptionResnetV2 , builds upon these previous works to achieve improved accuracy in emotion recognition tasks.

# **Proposed System**

The model proposed in this paper uses an ensembled learing approach which combines an EfficientnetB0 , Resnet50, InceptionResnetV2 and uses an MLP as a meta classifier to improve the accuracy of the model. Each model are first independently trained on the pre-processed input data. The outputs of each model are then sent as an input into an MLP , which acts as a meta classifier .This ensembling allows us to capture the strengths of individual models, resulting in improved accuracy and robustness.

## **Dataset and Preprocessing**

The FER 2013 dataset has been utilized for this project. It consists of grayscale images that represent various facial expressions that come under the 7 distinct emotions, namely anger, sadness, disgust, fear, happiness, surprise, and neutral. It contains close to 35000 images . Fig 1. Sample Images From Fer 2013 Dataset

To ensure uniformity, the input images are first resized to an input shape of 100x100 pixels. They are also normalized to a range of 0 to 1 to further improve the model’s efficiency. To align with the model’s output requirements, the labels are encoded. All of these combined helps in establishing a facial emotion recognition model with a solid foundation.

## **Model Architechture**

The proposed model contains 3 primary Convolutional Neural Networks, EfficientNetB0, InceptionResNetV2, and ResNet50. These models are then integrated through Multi-Layer Perceptron (MLP) which is being used as a meta classifier.

1. **EfficientNetB0:** Due to its efficient scaling capabilities, it works as the ensemble model’s backbone. A compound scaling technique is implemented in order to scale network depth, width and resolution uniformly. The model’s parameter count is reduced while simultaneously maintaining high accuracy by the use of depth wise separable convolutions. EfficientNetB0 leverages features from related tasks by pre-training it on a large dataset. This enhances the effectiveness of the FER 2013 dataset. A custom layer that contains a Global Average Pooling Layer and a dense layer is added. The dense layer is of 512 units. And it is followed up on by a SoftMax activation function for the 7-class output.
2. **InceptionResNetV2:** This model features inception blocks that capture multiscale representation, by integrating Inception architecture with residual learning. In order to recognize complex data patterns, this method is used. The vanishing gradient problem is mitigated and convergence is improved due to the addition of residual connection aids in training deeper networks. Same as EfficientNetB0, a custom layer that contains a Global Average Pooling Layer and a dense layer is added. The dense layer is of 512 units. And it is followed up on by a SoftMax activation function for the 7-class output.
3. **ResNet50:** The third model used in the project, ResNet50 helps in training deep architectures without degradation issues by employing residual connections to It. This model excels in complex image classification tasks such as facial emotion recognition by learning rich, hierarchical features. The custom top layer used in ResNet50 happens to be the same as the other models, consisting of a Global Average Pooling Layer and a dense layer is added. The dense layer is of 512 units. And it is followed up on by a SoftMax activation function for the 7-class output.

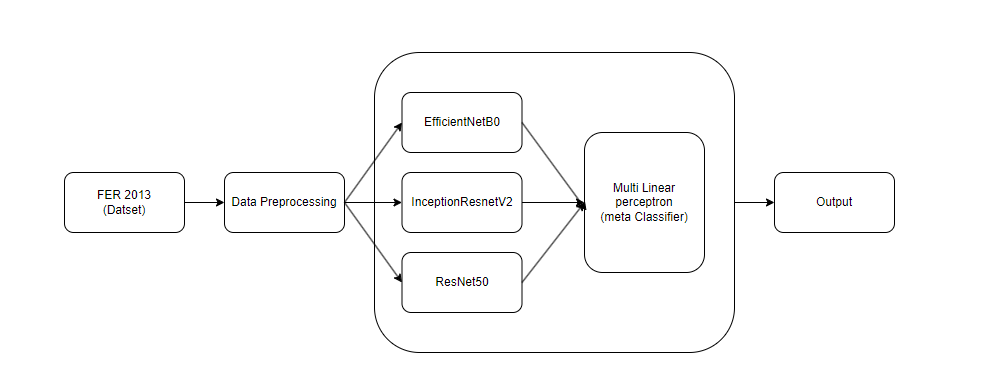


Fig 2 . Proposed System Architecture

1. **MLP:** Acts as an ensemble’s meta classifier, Multi-Layer Perceptron (MLP) is able to combine the outputs of EfficientNetB0, ResNet50 and IceptionResNetV2. The outputs of these models are complied into a feature matrix. The MLP is configured to have single hidden layer of 100 units and then it is trained on these matrices with true labels to optimize its weights. This technique allows the MLP model to learn complex relationships among the base model predictions in return improving the overall performance of the ensemble in emotion recognition tasks.

## **Model Training Process**

#### This training method follows a structured approach that promotes effective learining and generalization. Firstly, the three models EfficientNetB0, ResNet50 and IceptionResNetV2 were individually trained on the FER 2013 dataset, which contains imagaes of faces with labels of the emotions. The models were each trained for 50 epochs using a Stoochastic Gradient Descent (SGD) with Nesterov Momentum as the optimizer. Furthermore, to enhace the model performance, Reduce Learning Rate on Plateau callback was also applied, which was able to decrease learining rate when performance plateaued, hence allowing for more effective convergance.

#### Next, the outputs from these models were ensembled and used as the input features to the meta classifier, Multi-Layer Perceptron (MLP). The MLP is able to find and learn the relationships between the models and also accute to the diiferences between predicted and true class labels. The MLP was made to execute for up to 300 iterations, while enforcing convergence to a stable solution.

Finally, the ensembled model was evaluated for the model’s performance by using a separate test dataset, with metrics such as accuracy, precision, recall and F1-score which were used to assess its effectiveness in the task of emotion recognition. This method of ensemble learning was able to leverage the strengths of each model while being able to minimize their individual limitations, returning a model with highly accurate emotion recognition system.

# Experimental Results and Discussion

The proposed model TriceptionNet’s performance was evaluated using the FER 2013 dataset. The model which comprises of three individual architectures: EfficientNetB0, ResNet50 and InceptionResNetV2 each were individually trained for 50 epochs. The training process involved the use of Stochastic Gradient Descent (SGD) with Nesterov Momentum and also utilized a learning rate reduction strategy upon receiving a plateau. The below fig 3, fig 4 and fig 5 indicate the training and validation accuracies of each model over the 50 epochs.

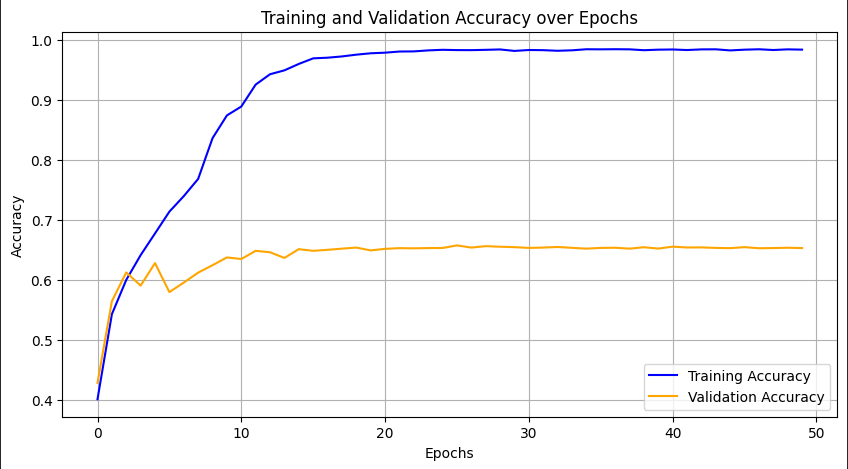


Fig 3. Training And Validation Accuracy of EfficientNetB0

A graph with blue and orange lines

Description automatically generated

Fig 4. Training And Validation Accuracy of InceptionResNetV2 over 50

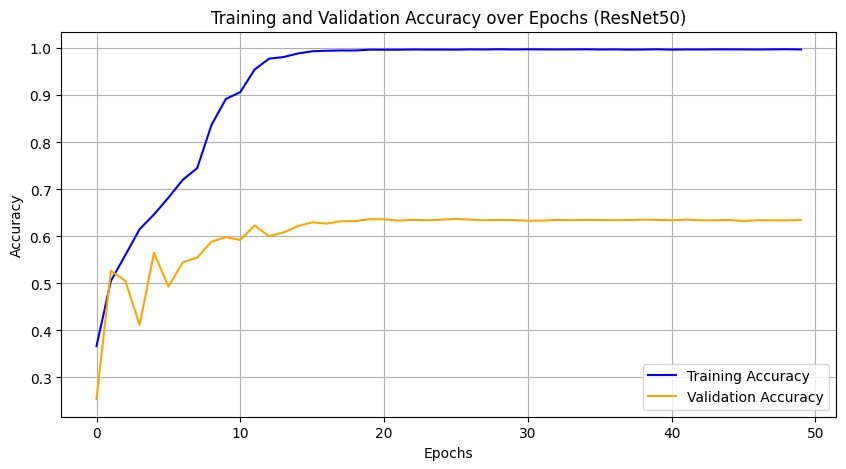


Fig 5. Training And Validation Accuracy of Resnet50 over 50

The results of test done on each model are summarized in **Table 1**, which is able to provide better insights into their individual performance metrics, including the test loss and accuracy. Furthermore, the results of the tests done on the ensemble model using Multi-Layer Perceptron (MLP) is given in **Table 2**, which elaborates precision, recall and F1-score for each class of emotion.

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| EfficientNetB0 | 0.6524 |
| InceptionResNetV2 | 0.6169 |
| ResNet50 | 0.6304 |

Table 1. Independent Model Accuracies

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| 0 | 0.80 | 0.77 | 0.79 | 831 |
| 1 | 0.55 | 0.48 | 0.52 | 1024 |
| 2 | 0.61 | 0.58 | 0.59 | 958 |
| 3 | 0.63 | 0.62 | 0.62 | 1233 |
| 4 | 0.52 | 0.59 | 0.55 | 1247 |
| 5 | 0.83 | 0.54 | 0.66 | 111 |
| 6 | 0.83 | 0.88 | 0.86 | 1774 |
| Macro Average | 0.68 | 0.64 | 0.65 | 7178 |
| Weighted Average | 0.67 | 0.67 | 0.67 | 7178 |

Table 2. Performance Metrics of Ensemble MLP

The results from each model individually indicates a distinct variance in performance, with the model EfficientNetB0 providing the highest test accuracy of 0.6524, followed by ResNet50 at 0.6304 and InceptionResNetV2 with 0.6169. These performances suggest that each model demonstrates unique strength, but no single model excels across the emotion classification task. This variance implies the importance of model selection based on the requirements and tasks, as each architecture may be able to capture different aspects of the dataset.

Compared to the individual models the ensembled model has significantly outperformed each of them, achieving an overall accuracy of 0.67. By combining the outputs of EfficientNetB0, ResNet50 and InceptionResNetV2 by using a Multi-Layer Perceptron (MLP) as the meta-classifier, the ensemble technique was able to utilize the strengths of each model resulting for a more robust model. This method was able to reduce the weaknesses of the inherent models, more identifiable in classes with fewer samples, which is evident by the improved precision and recall values. The improved performance of the ensembled model shows that leveraging the strategy of ensemble learning for multiple models can lead to enhanced accuracy and better classification in facial recognition task.

# Conclusion

The study introduces an approach to facial emotion recognition through the use of ensemble learning of multiple architectures, specifically EfficientNetB0, ResNet50 and InceptionResNetV2. These models were trained for 50 epochs independently, highlighting their unique strengths with the EfficientNetB0 achieving the highest accuracy individually. The true potential of the model was realized through the ensemble method, resulting in an overall accuracy of 0.67 showcasing its potential over standalone models.

These finding helps understand the advantages of model ensembling for enhancing classification performance and also for complex task like facial emotion recognition. By utilizing the different strengths of these architectures, the ensemble method not only improved the accuracy but also the consistency over different emotion classes. Future works in this could explore different ensemble strategies and additional hybrid architectures which could further enhance the accuracies. Overall this work provides with precious insights into the techniques for emotion recognition while paving the paths for advanced applications within the fields of facial emotion classification and human-computer interaction.

##### References

1. Kahou, S. E., Bouthillier, X., Lamblin, P., Gourier, C., & Bengio, Y. (2017). Far above average: A deep learning framework for facial expression recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2829-2837
2. Zhang, Y., Zhang, Y., Li, Z., & Qiao, Y. (2018). Ensemble deep learning for facial expression recognition. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6
3. Li, X., Zhang, S., & Zhang, Y. (2019). Attention-based convolutional neural network for facial expression recognition. IEEE Transactions on Affective Computing, 10(2), 232-242.
4. Gao, Y., Wang, Y., & Liu, X. (2020). Facial expression recognition using deep learning with attention mechanism and temporal information. IEEE Access, 8, 156869-156878.
5. Guo, Y., Li, Y., & Wang, C. (2021). Data augmentation for facial expression recognition with deep learning. IEEE Access, 9, 113682-113692..
6. Mollahosseini, A., Chan, M., & Mahoor, M. H. (2017). Deep learning for facial expression recognition: A survey. IEEE Transactions on Affective Computing, 8(1), 18-39.
7. Kim, J., Kim, J., & Lee, J. (2022). Ensemble learning for facial expression recognition with deep convolutional neural networks. IEEE Access, 10, 109127-109136..

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**