**CSCI 323 Modern AI: Group Assignment**

**Group 2: Facial Expression Recognition**

An analysis on the Facial Expression Recognition (FER) problem: Comparison of complexity of models on accuracy in FER

[**Introduction**](#_xmolle1rp7ij) **2**

[**Background Theory**](#_6cgng6xbsaoy) **3**

[**Solutions, Evaluation, and Discussions**](#_m52f0xlmfxpe) **4**

[**Conclusion**](#_hjxrs5t1awjo) **7**

[**References**](#_xs8noskr5bvq) **8**

| **Name** | **UOW ID** | **Contribution** |
| --- | --- | --- |
| Lai Wei Jie | (UOW: 7770042 | **100%** |
| Farah Nurliyana Bte Roseley | (UOW: 8377443) | **100%** |
| Wasilah Mumina | (UOW: 8434463) | **100%** |
| Ivan Yong Sheng, Tan | (UOW: 8750099) | **100%** |
| Nila Kandasamy Rajamohan | (UOW: 8433252) | **100%** |
| Ray Mar Jun Kang | (UOW: 7769829) | **100%** |

# 

# Introduction

Facial expressions are crucial in communication and convey complex mental states during interaction. In non-verbal communication, the face transmits emotions ([Darwin and Prodger, 1996](https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1359471/full#B6)).

“Facial Expression Recognition (FER) is a computer vision task aimed at identifying and categorizing emotional expressions depicted on a human face.” (Facial Expression Recognition, n.d.). This is done by automating the process of, analyzing human facial features, and mapping them to emotions.

FER is increasingly becoming a vital tool in various fields, offering enhanced capabilities in understanding human emotions through technology. The automation of this process, particularly through advanced machine learning techniques, provides not only efficiency but also the ability to analyze vast amounts of data that would be infeasible manually.

Use cases of FER might vary, but are always closely related to Sentiment Analysis.

Sentiment Analysis might allow better understanding and a gauge of a customer’s reactions to a company’s product or marketing strategy, by reading emotions of customers interacting with brands and products through a website or in real life. This might be a great way to measure consumer engagement with a product.

Another use case of Sentiment Analysis, especially in a time of Interactive or Affective computing and Large Language Models, is the development of a system with the simulation of empathy. Affective computing is an interdisciplinary field that involves studying and developing systems capable of understanding and interpreting human emotions ([Banafa, 2016](https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1359471/full#B2)).

Moreover, FER's integration into artificial intelligence systems has broader implications, particularly in fields like mental health, where emotion detection can assist in diagnosing conditions or monitoring patient progress. Additionally, in security, FER can be used where emotion recognition can enhance surveillance systems by identifying individuals who may pose a threat based on their emotional states.

A system capable of delivering empathetic responses, greatly enhances the user’s experience, and the company that is able to create such a system will be able to capture a majority share of the market, in a winner takes all scenario.

In this project, we aim to leverage the FER2013 dataset to train and evaluate various CNN architectures, systematically increasing their complexity to observe the impact on accuracy and performance. By analyzing these models, we hope to contribute to the understanding of how neural networks can be optimized for emotion recognition tasks, ultimately paving the way for more sophisticated and reliable FER systems.

# Background Theory

The common theme behind any Machine Learning project is the quality and size of the dataset, being used to train the models. In the context of FER, these factors are particularly crucial, as they directly impact the model's ability to generalize across different facial expressions and diverse populations.

FER-2013

The FER-2013 dataset, which was first created for an ongoing project by Pierre-Luc Carrier and Aaron Courville, was introduced during the ICML 2013 Challenges in Representation Learning. It has since become one of the most widely utilized datasets for facial expression recognition tasks, providing a benchmark for developing and evaluating models in this field.

The FER-2013 dataset consists of 35,887 grayscale images, each with a resolution of 48x48 pixels. These images are categorized into seven emotion classes (except contempt):

**0:** Angry: 4,593 images**1:** Disgust: 547 images**2:** Fear: 5,121 images**3:** Happy: 8,989 images**4:** Sad: 6,077 images**5:** Surprise: 4,002 images**6:** Neutral: 6,198 images

The FER-2013 dataset was curated using Google Image Search, a method that introduces potential biases due to the non-random selection of images. This approach also results in varying image quality, with some examples in the dataset (shown below) being notably poor in terms of resolution and relevance.



**Dataset Criticisms and Corrective Strategies**

“Facial expression data is usually highly skewed. This form of imbalance is commonly referred to as intrinsic variation, i.e., it is a direct result of the nature of expressions in the real world.” (Mollahosseini, Hasani, & Mahoor, 2017, p. X).

It is to be expected that there will be a vast imbalance of sample sizes, and some preprocessing required before model training, either through synthetic generation, over-sampling and data augmentation of undersampled categories.

Data augmentation is a strategy of creating samples, by either Geometric Transformation (rotate, translation, flipping), Color Transformation (Brightness, Contrast), Noise Injection and Cropping. This helps to improve the generalization ability of the model, making it more robust to variations in real-world data.

# Solutions, Evaluation, and Discussions

Notebook: <https://colab.research.google.com/drive/1s7Q4iiFTPuiROh93CpQaLJ8sMpW-iF0j>

**The goal of this project is to demonstrate how using increasingly complex CNN models, with the same variables, affect overall accuracy.**

**The baseline accuracy to meet is 60.7%, which represents the average human accuracy for emotion recognition on the AffectNet dataset.**

In this study, we evaluate three baseline models:

1. Simple CNN
2. ResNet18 via Transfer Learning, a more complex CNN architecture
3. DenseNet 121 via Transfer Learning, a more complicated and deeply connected architecture

Initially, the VGG16 model was considered for transfer learning due to its proven effectiveness in image classification tasks. However, the extensive training time required on Google Colab exceeded our resource limits, leading us to select the more computationally efficient ResNet18 model instead.

Given the strong performance of CNNs in image recognition tasks, we will attempt to train a simple CNN as a baseline, a more advanced architecture of CNN like ResNet18 through transfer learning, and another baseline model like Support Vector Machines.

ResNet18 is a variant of the ResNet (Residual Network) architecture introduced by Kaiming He et al. in their 2015 paper “Deep Residual Learning for Image Recognition.” ResNet18 is known for its simplicity and effectiveness, providing a good balance between model complexity and performance. This architecture includes 18 layers composed of convolutional, pooling, and fully connected layers. ResNet18 was trained on approximately 1.2 million training images from the ImageNet dataset. The images span 1,000 different categories, including animals, objects, scenes, and more.

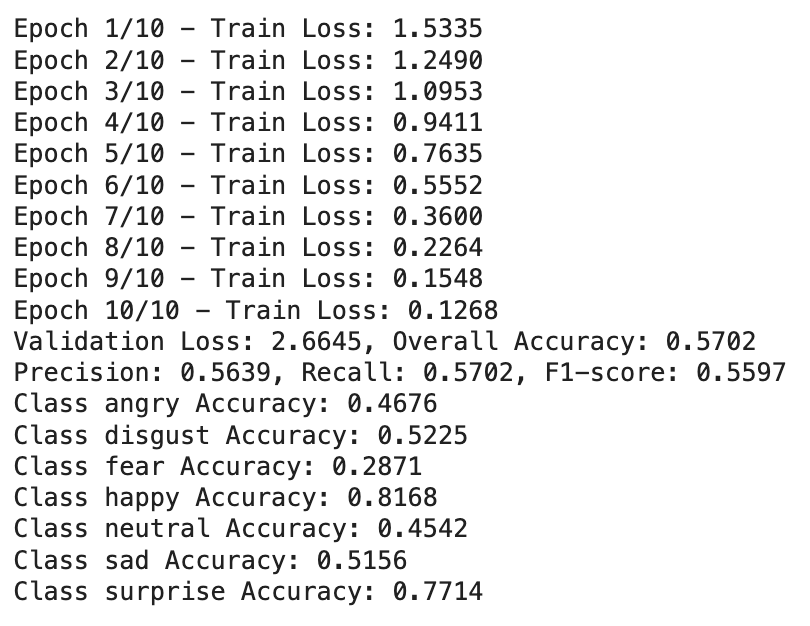
Our third model: DenseNet121 consists of 121 layers where each layer is connected to every other layer in a feed-forward fashion.

This means that the feature maps of all preceding layers are used as inputs for each subsequent layer. These dense connections encourage feature reuse throughout the network and mitigate the vanishing-gradient problem. Transition layers also perform compression to reduce the number of feature maps, which helps in controlling the model's complexity and computational requirements.

**FER2013 Dataset obtained from:** <https://www.kaggle.com/datasets/msambare/fer2013>

1. **Simple CNN Model Performance (FER2013)**

* **Architecture**: A simple CNN consisting of three convolutional layers, each followed by ReLU activation functions and max-pooling layers. Following the convolutional layers, the model includes two fully connected layers, which serve as the classifier.
* **Loss Function**: Cross-entropy loss, suitable for multi-class classification problems. It is particularly effective in scenarios where the classes are mutually exclusive, as it penalizes incorrect predictions more harshly, thus driving the model to learn more discriminative features for each class.
* **Optimizer**: Adam optimizer with a learning rate of 0.001, which is widely used for training deep learning models due to its efficiency and good performance.

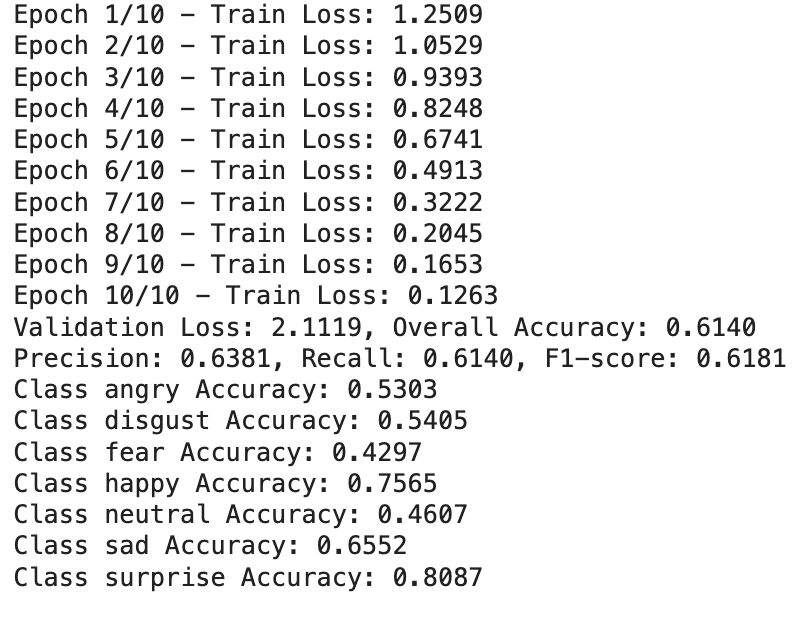


**Overall Accuracy: 57%**

*While this performance is below the human benchmark of 60.7%, it serves as a crucial baseline for evaluating the impact of more complex CNN architectures. The relatively low accuracy highlights the limitations of simple models in handling the intricacies of facial expression recognition, especially when dealing with subtle differences between emotions.*

1. **ResNet18 Model CNN Model Performance (FER2013)**

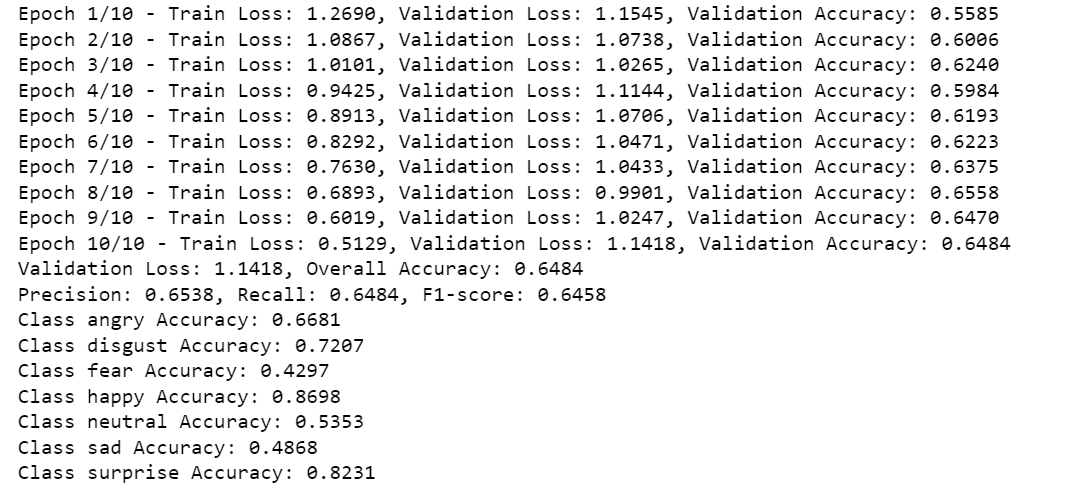
* **Architecture**: A more complex CNN featuring an input layer with 3 channels, followed by 4 residual blocks. Each block contains two 3x3 convolutional layers, and the model concludes with a fully connected layer that outputs the final classification.
* **Loss Function**: Cross-entropy loss, suitable for multi-class classification problems.
* **Benefits:**  Utilizes deeper layers and residual connections to capture more intricate details in facial expressions, resulting in more accurate classifications compared to simpler models.

****

**Overall Accuracy: 61.4%**

1. **DenseNet121 (FER2013)**

* **Architecture:** DenseNet121 is a highly complex CNN with 121 layers, characterized by dense connections between layers. Each layer receives input from all preceding layers, enhancing feature reuse and ensuring better gradient flow throughout the network. The architecture includes:
  + **Input Layer:** Accepts images with 3 channels (RGB).
  + **Dense Blocks:** Four dense blocks, where each block contains multiple layers, each connected to every other layer within the block. These dense blocks are interleaved with transition layers that include convolution and pooling operations to manage the size of the feature maps.
  + **Fully Connected Layer**: Ends with a fully connected layer (classifier) that performs the final classification.
* **Benefits:** The dense connections in DenseNet121 significantly improve information flow and gradient propagation, addressing common challenges like the vanishing gradient problem. By reusing features from previous layers, DenseNet121 enhances learning efficiency, allowing the model to capture more complex and abstract features with fewer parameters compared to traditional CNNs.

****

**Overall Accuracy: 64.8%**

# Conclusion

Training a model to effectively solve the Facial Expression Recognition (FER) problem is a complex and challenging endeavor, especially when attempting to surpass human-level performance. Even the most advanced published models often struggle to achieve very high levels of accuracy, typically in the range of 90% or above, which is often necessary in scenarios where the consequences of misclassification are severe, such as in legal or security-related applications.

Despite significant advances in deep learning, the challenge of accurately classifying emotions from facial expressions remains. Several factors contribute to this difficulty:

* **Annotation Challenges:** Discrepancies in labeling datasets can lead to inconsistencies in model training, resulting in reduced accuracy.
* **Dataset Quality:** The inclusion of non-face images or low-quality images with poor contrast and sharpness further complicates the training process.
* Data Collection Limitations: Capturing a diverse and representative dataset from natural environments remains a hurdle, impacting the model's ability to generalize.
* **Resource Constraints:** The computational power required to train sophisticated models is often prohibitive, particularly when dealing with large datasets and complex architectures.

In our study, the ResNet18 model achieved an accuracy of 61.4%, just surpassing the average human performance benchmark of 60.7%. However, the DenseNet121 model further improved accuracy to 64.8%, demonstrating that more complex and optimized models can yield better performance. These results highlight the potential benefits of using deeper architectures that can capture more abstract features necessary for distinguishing subtle differences in facial expressions.

However, there are other strategies available to boost the performance to meet the benchmarks required, instead of simply just using more complex models. In “Papers with Code”, the top performing model on the FER2013 dataset is an Ensemble ResMaskingNet (ResNet variant) with 6 other CNN’s, with an accuracy of 76.82%. The main differential is the ensemble method, in conjunction with other CNN models. This suggests that leveraging the strengths of multiple models through ensemble techniques can significantly enhance accuracy beyond what individual models can achieve.

Another critical challenge in FER is the class imbalance within the dataset, which leads to poor prediction accuracy for certain emotions, particularly those less represented, such as fear and sadness. Addressing this issue requires implementing strategies like oversampling, synthetic creation of underrepresented classes, or adjusting the class weights during training. Such methods can help ensure that the model is not biased towards more frequent emotions, thereby improving its overall robustness and fairness.

In conclusion, while the current state of FER models shows promise, they are best utilized as supplementary tools rather than standalone solutions. Their effectiveness can be enhanced by integrating them with other evaluation metrics and methodologies, ensuring a more comprehensive

In its current state, FER models should best be used as a supplementary tool, in addition to other evaluation metrics to gauge sentiment and emotion.

# References

Facial Expression Recognition. (n.d.). *Papers with Code*. https://paperswithcode.com/task/facial-expression-recognition#task-home

Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor, “AffectNet: A New Database for Facial Expression, Valence, and Arousal Computation in the Wild”, IEEE Transactions on Affective Computing, 2017.

Challenges in Representation Learning: A report on three machine learning contests, Ian Goodfellow et al., 2013

Neurodivergent Insights. (n.d.). Arousal Valence Model. *Neurodivergent Insights*. https://neurodivergentinsights.com/blog/arousal-valence-model