Comparison of Facial Expression Classification Methods

COMP 6731 - Pattern Recognition

Fall 2023

Submitted To: Adam Krzyzak

Team Members:

Name	Student ID
Gurleen Pannu	40289625
Saketh Oppula	40221013
Thuhin Khanna Rajesh Kannan	40271685

GitHub: https://github.com/sakethoppula/COMP_6731_project

TABLE OF CONTENTS

1. ABSTRACT

Facial expression recognition plays a crucial role in the field of computer vision and pattern recognition, providing a distinctive approach to interpreting human emotions through digital methods. This project involves a thorough examination of different model architectures and preprocessing methods tailored for the classification of facial expressions. The main objective is to create and deploy a proficient pattern recognition model, specifically utilizing various Convolutional Neural Network (CNN) types. Additionally, the project aims to meticulously compare the model's performance across diverse datasets, including CK+, JAFFE, and BU-3DFE repositories. These repositories offer a wide range of facial expressions, encompassing emotions like anger, disgust, fear, happiness, sadness, and surprise.

In addition to assessing the model's performance, the project emphasizes the importance of investigating methods for preprocessing facial data originating from diverse ethnicities, facial structures, and orientations. The preprocessing phase will meticulously tackle noise introduced by variations in contrast, shadows, and other environmental factors. The research acknowledges the critical role of effective preprocessing in enhancing the model's robustness and generalizability.

The implementation will leverage the capabilities of Convolutional Neural Networks, a category of deep learning models ideally suited for tasks related to images. The selection of CK+, JAFFE, and BU-3DFE datasets is based on their public accessibility and comprehensive annotations of facial expression labels, establishing a strong basis for the training and evaluation of the model.

Furthermore, the project situates itself within the Python programming domain, making use of pre-existing machine-learning libraries to simplify the implementation procedure. The effectiveness of the classification model will be thoroughly assessed, comparing its performance against established models specifically designed for similar tasks within the selected dataset. The input pipeline involves managing facial expression images in .png and .tiff formats, requiring preprocessing measures to eliminate potential noise and standardize the input size.

Ultimately, the classification model will produce confidence values linked to each pre-defined emotion class. This endeavour aims not only to add to the expanding realm of research in facial expression recognition but also to offer insights into the effectiveness of convolutional neural networks when juxtaposed with current models. In doing so, it strives to advance the comprehension and application of emotion recognition technology.

2. INTRODUCTION

● Facial Expression Recognition (FER)

Facial Expression Recognition (FER) is a technique that uses computer algorithms to recognize and interpret human facial emotions. This area has become more well-known in the fields of computer vision and artificial intelligence because of its many uses, which include emotion-aware systems and human-computer interface (HCI). FER's main goal is to make it possible for computers to decipher and comprehend human emotions via visual clues such as raised eyebrows, frowns, and grins.

The first step in the FER process is usually the collection of face data via picture or video input. Then, sophisticated computer vision methods—which frequently make use of deep learning models—are used to identify and evaluate face characteristics like the mouth, nose, and eyes. These characteristics are essential for deriving pertinent face expression data. In order to identify patterns and correlations between certain face configurations and associated emotions, machine learning algorithms are trained on datasets including labelled facial expressions. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are two popular methods that are particularly good at capturing the temporal and spatial correlations seen in facial data.

FER is used in many different fields. In HCI, it makes it possible for systems to react to users' emotional states in real time, improving the user experience as a whole. It helps to create more realistic and engaging virtual reality experiences by adjusting virtual settings according to users' emotional reactions. Furthermore, emotion-aware machines may be applied to marketing, healthcare, and other domains to analyze emotional responses and yield insightful data for decision-making. In the subject of FER, research is still ongoing despite notable advancements in addressing issues including managing a variety of facial expressions and guaranteeing resilience in real-world scenarios.

● Convolutional Neural Networks (CNN)

Since convolutional neural networks (CNNs) are so good at identifying patterns and spatial hierarchies in image data, they have shown to be very useful in the field of facial expression recognition. CNNs are a class of deep learning models that are well-suited for facial expression analysis because they were created especially for image processing tasks. Convolutional layers, which are part of CNN architecture, are designed to systematically scan and extract features from input images. These layers are followed by fully connected layers that categorize the extracted features and pooling layers that decrease spatial dimensions while keeping pertinent information.

CNNs are particularly good at automatically extracting discriminative facial features that are necessary for emotion recognition in the context of FER. CNNs are trained to identify patterns linked to various facial expressions, including raised eyebrows, frowns, and smiles. In order to detect hierarchical features like edges, textures, and facial landmarks, the convolutional layers function as learnable filters. The network can effectively generalize to a variety of facial characteristics thanks to this hierarchical feature extraction, which is essential for capturing the subtle details of facial expressions.

The accuracy and robustness of FER have significantly improved as a result of the use of CNNs. Effective transfer learning is made possible by the ability to refine pre-trained CNN models, like VGGNet, ResNet, or Inception, on FER datasets. Even in situations where there is a shortage of training data, this transfer learning strategy makes use of the insights gleaned from massive image datasets to improve the performance of FER models. In order to achieve state-of-the-art results in facial expression recognition, CNNs and transfer learning have been combined. This has made it possible to interpret and comprehend human emotions through facial cues with great power.

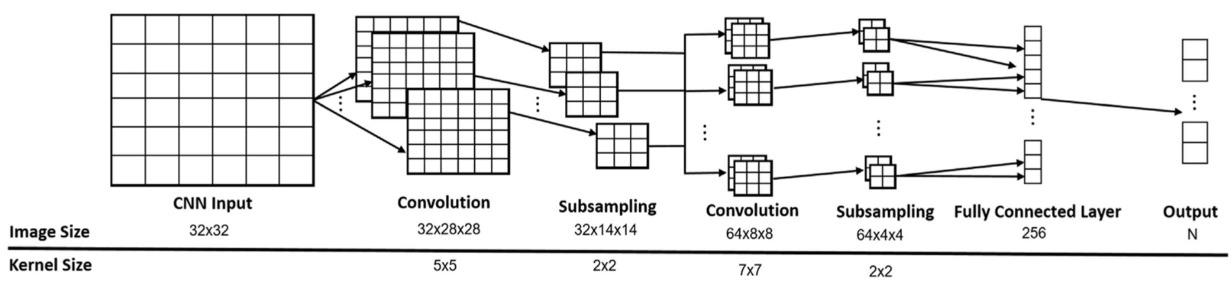

Fig. 1 - CNN Model for classifying facial expression having 3 convolutional layers and 2 sub-sampling layers with ReLU.

- ## Residual Networks (ResNet)

Residual Networks, or ResNets for short, are game-changing developments in deep learning that provide an answer to the problems associated with training extraordinarily deep neural networks. The addition of residual connections, which enable the network to learn residual functions—capturing the variation between an input and an output of a layer—is what makes them unique. Within the domain of Facial Expression Recognition (FER), ResNets demonstrate their mastery by skillfully identifying minute facial cues linked to a range of emotions. The architecture's success in extracting hierarchical features from facial data that are relevant to comprehending complex emotional states can be attributed in large part to its unique ability to facilitate the training of very deep models.

Transfer learning methodologies are frequently employed in FER applications that utilize ResNets. The base is provided by pre-trained ResNet models on large-scale image datasets like ImageNet. After pre-training, these models are refined on more specialized FER datasets, utilizing the insights gathered. ResNets are an invaluable tool for FER tasks because of their versatility and effectiveness in generalizing across various datasets, especially in situations where labelled facial expression data is scarce. Skip connections are incorporated into ResNets to facilitate smooth information flow and help preserve important details about facial expressions during training. As a result, ResNets are now a key component in pushing FER's accuracy limits and influencing developments in a variety of fields, including emotion-aware computing and human-computer interaction.

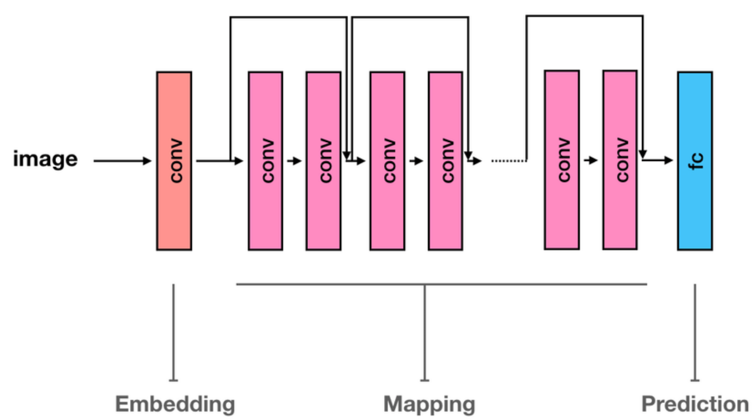

Fig. 2 - A schematic view of ResNet architecture, decomposed into 3 blocks: Embedding, Mapping and Prediction.

3. OVERVIEW

The basis of the work done is the project draws from the work of A.T. Lopes et. al., "Facial Expression Recognition with Convolutional Neural Networks: Coping with few data and the training sample order" [1]. Their work deals with the problem of recognising facial expressions from the CK+, JAFFE and BU-3DFE datasets. The work identifies seven basic emotions namely anger, contempt, disgust, fear, happiness, surprise and sadness. Using a CNN classifier and preprocessing the images prior to training they were able to classify the facial expressions with a high accuracy of 96.76%. The preprocessing techniques used in the paper are chosen specifically to increase the efficiency of the classifier. One of the preprocessing techniques is image cropping which reduces the problem area so that the model is not exposed to unwanted data that is in the background of the images.

The work also talks about synthetic data generation as a means of coping with fewer data samples. Rotating the images can be a way of increasing the number the data samples available. This makes the model learn more general patterns and adapt to any new data that it may encounter in the future. The project builds on the work by exploring other processing techniques that can be used in place of the ones from the original work or in tandem with them. Drawing comparisons between different processing techniques based on accuracy measures can expose details about the nature of the data that the model uses as a deciding factor in determining the expression found in the images.

4. RELATED WORKS

A. Facial Expression Recognition using Convolutional Neural Networks: State of the Art

The paper [2] discusses the significance of automatic facial expression recognition in human-computer interaction and various applications. It notes the active research in this area, particularly focusing on CNNs for feature extraction and inference. The paper aims to review the state of the art in image-based FER using CNNs, highlighting algorithmic differences and their impact on performance. The paper emphasizes the importance of FER in nonverbal communication and its applications. While recognizing basic expressions in controlled conditions is considered a solved problem, recognizing expressions under naturalistic conditions remains challenging. CNNs are seen as a potential solution due to their success in related tasks.

The paper reviews six state-of-the-art methods for CNN-based FER, discussing their methodologies and performances, particularly on the FER2013 dataset. The dataset is described as challenging due to variations in age, pose, and other factors. The overview highlights different approaches, including ensemble methods, committee models, and the use of diverse CNN architectures. Methodological differences among the reviewed works are discussed in terms of preprocessing, CNN architecture, and training/inference. Preprocessing involves face detection, registration, and illumination correction. CNN architectures vary significantly in depth and parameters, with a focus on shallower networks. Training and inference methods differ, with some using data augmentation and ensemble techniques.

In conclusion, the paper identifies existing bottlenecks in CNN-based FER and suggests directions for improvement. Overcoming one bottleneck, the relatively basic architecture of CNNs is shown to significantly improve performance. An ensemble of modern deep CNNs achieves a test accuracy of 75.2% on FER2013, outperforming existing methods.

The paper concludes by expressing the expectation that addressing remaining bottlenecks will lead to further performance improvements and mentions future research directions, including data augmentation and the creation of a more comprehensive FER dataset.

B. Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields

The paper [3] proposes a two-part network for Spatio-Temporal Facial Expression Recognition using Convolutional Neural Networks (CNNs) and Conditional Random Fields (CRFs). The existing methods often rely on hand-crafted features and traditional classifiers, requiring rigorous tuning. In contrast, the proposed approach combines a CNN-based architecture with a CRF module for recognizing facial expressions in videos.

The first part of the network captures spatial relations within facial images using convolutional layers, InceptionResNet modules, and fully-connected layers. The second part incorporates a linear chain CRF to capture temporal relations between image frames in videos. The paper evaluates the proposed network on three publicly available databases (CK+, MMI, and FERA) in subject-independent and cross-database experiments. The results show that combining the deep network with the CRF module significantly improves facial expression recognition in videos, outperforming state-of-the-art methods in cross-database experiments and achieving comparable results in subject-independent experiments.

The experimental setup involves preprocessing steps, resizing images, and using 5-fold cross-validation. The proposed network outperforms existing methods in recognizing facial expressions across different databases, demonstrating the effectiveness of the combined spatial and temporal modelling. The authors suggest future work involving the evaluation of different CRF variations and residual connections in the network to enhance feature maps for facial expressions.

C. Facial Expression Recognition with Temporal Modeling of Shapes

The research paper [4] introduces an innovative framework for automatic facial expression recognition in video sequences, focusing on the temporal dynamics of facial shapes. Utilizing Latent-Dynamic Conditional Random Fields (LDCRFs), the proposed method surpasses traditional models like Conditional Random Fields (CRFs) and Support Vector Machines (SVMs). The study addresses the challenge of recognizing facial expressions with brief and subtle movements by emphasizing the temporal variations in facial shapes, crucial for discriminating between similar expressions. Shape features, represented by landmark points and analyzed through Generalized Procrustes Analysis and Principal Component Analysis (PCA), are demonstrated to be more informative than appearance features alone.

Experiments conducted on the Extended Cohn-Kanade Dataset (CK+) provide empirical evidence of the method's efficacy, especially in differentiating between nuanced expressions and the neutral state. The results highlight the superior performance of the LDCRF approach, particularly for challenging expressions like sadness and anger. The paper concludes by underscoring the importance of effectively modelling small facial motions for accurate facial expression recognition, suggesting potential extensions to broader datasets and real-time video applications, and indicating the broader implications of the proposed methodology in advancing the field of facial expression recognition.

5. DATASET DESCRIPTION

The CK+ (Cohn-Kanade) dataset is a widely used and well-known benchmark in the field of Facial Emotion Recognition (FER). It was developed by researchers at the University of Pittsburgh and the Kanade Research Group. The CK+ dataset is specifically designed for studying facial expressions and contains a comprehensive collection of posed facial expressions of basic emotions [5].

The dataset was created by capturing the images of 210 adults with age between 18 and 50 years old, spanning a diverse range of ethnicities, ages, and gender. The subjects in the dataset are 69% female, 81% Euro-American, 13% Afro-American and 6% of other groups.

The image sequences for the frontal views and the 30-degree views were digitized into 640x480 pixel arrays with 8-bit grey-scale or 24-bit colour values.

The dataset consists of each participant posing 8 expressions (as shown in Fig. 2), which are as follows:

- Anger
- Contempt
- Disgust
- Fear
- Happy
- Surprise

- Sadness
- Neutral

Fig. 3 – Facial Expressions captured in the CK+ dataset

6. METHODOLOGY

● Pre-processing Techniques

When it comes to improving the quality and effectiveness of image datasets for a range of machine learning applications, data preprocessing is extremely essential. The preprocessing of data is crucial before incorporating it into a model, as image databases frequently have intricate and varied information. Data preprocessing is important because it can reduce overfitting, improve model performance, and make sure that the model can extract relevant patterns from the images. All of these things can lead to machine learning models that are more reliable and accurate when used for image analysis tasks.

The paper proposes a simple solution to Facial Expression Recognition which includes a combination of Convolutional Neural Network (CNN) and specific image preprocessing steps that include – Rotation Correction, Image Cropping, Resizing the Image and Intensity Normalization. Our goal in this project is to determine how the different preprocessing techniques on the dataset affect the accuracy of the model (training accuracy, validation accuracy and testing accuracy).

For this purpose, we will be pre-processing the dataset using the following pre-processing techniques:

- **Conversion to grayscale**

  It is a fundamental preprocessing technique which involves transforming colour images into grayscale, where each pixel is represented by a single intensity value, eliminating colour information. By doing so, the computational complexity is reduced, as grayscale images contain only one channel instead of three (red, green, and blue in RGB). This not only saves memory and processing power but also helps in standardizing input data for models.

  Grayscale images often retain crucial structural information, making them suitable for tasks such as edge detection and texture analysis. The simplicity introduced by converting to grayscale can enhance model training efficiency, especially when colour information may not be crucial for the specific task at hand, making it a widely adopted preprocessing step in various image-based applications.

- **Noise Reduction with Gaussian Blurring**

  Noise reduction with Gaussian blurring is employed to enhance the quality of image datasets. Gaussian blurring involves convolving an image with a Gaussian filter, where each pixel's value is replaced by a weighted average of its neighbours, with the weights determined by a Gaussian function. This technique effectively smoothens the image by reducing high-frequency noise and small-scale details, contributing to a cleaner and more generalized representation.

  By suppressing noise, Gaussian blurring aids in improving the robustness of computer vision models, making them less sensitive to minor fluctuations in pixel values. This preprocessing step is particularly useful in scenarios where the input images may suffer from artefacts or random variations, helping models focus on essential features and patterns while minimizing the impact of irrelevant pixel-level fluctuations.

- **Face Detection**

  Face detection using the Haar cascade is a popular preprocessing technique for real-time applications. This technique relies on a Haar cascade classifier, which is used to identify objects in images or video. The Haar cascade is trained with positive and negative samples of faces to create a set of rules that define facial features. During face detection, the classifier slides over the image at different scales and locations, applying these rules to identify regions resembling faces.

The Haar cascade method is computationally efficient, making it suitable for real-time applications, such as video surveillance and facial recognition systems. Its ability to quickly eliminate non-face regions reduces the computational load, making it an effective preprocessing step for applications where speed and accuracy are crucial, providing a foundation for subsequent facial analysis or recognition tasks.

- **Normalization**

It is applied to image datasets to standardize and scale pixel values, ensuring uniformity across images. This technique involves adjusting the intensity values of pixels to a common scale, typically between 0 and 1 or -1 and 1. Normalization helps mitigate the impact of varying intensity ranges among images, ensuring that models are not unduly influenced by the absolute pixel values.

By normalizing the data, issues related to differing brightness or contrast levels are addressed, making the training process more robust and facilitating convergence in the models. Normalization is particularly vital when using deep learning frameworks, as it aids in stabilizing the training process and accelerates convergence by ensuring that the model's weights are updated consistently across diverse image inputs.

- **Histogram Equalization**

Histogram equalization is used to enhance the contrast and visibility of details. This method works by redistributing the intensity values of pixels in an image, making the histogram more uniform. By equalizing the histogram, the full dynamic range of pixel intensities is utilised, leading to improved visual quality and highlighting finer details in both dark and bright regions of the image. Histogram equalization is particularly beneficial when dealing with images with uneven lighting conditions or poor contrast.

This technique is commonly applied in computer vision and image processing tasks to standardize image appearance, making it easier for algorithms to detect features and patterns. While histogram equalization can be powerful in enhancing image quality, care should be taken to ensure that it is applied appropriately, as it may amplify noise in certain situations.

- **Edge Detection**

In our project, we use the Canny Edge Detector for the purpose of Edge Detection. Canny Edge Detection is widely used for identifying and highlighting edges within an image. The Canny Edge Detection algorithm involves multiple steps, including gradient computation, non-maximum suppression, and edge tracking by hysteresis. By detecting abrupt changes in pixel intensity, Canny Edge Detection effectively isolates the boundaries between objects or regions in an image, providing a clear representation of the underlying structures.

This technique is valuable in various computer vision applications, such as image segmentation, object recognition, and feature extraction. Canny Edge Detection is known for its ability to produce well-defined and accurate edges while minimizing the impact of noise. It plays a crucial role in enhancing the performance of subsequent image analysis tasks by emphasizing the most relevant features and details within the dataset.

Fig. 4 – Result of the different pre-processing techniques applied to the CK+ dataset

7. RESULTS & COMPARISON

The model was trained on the CK+ dataset using the k-fold cross validation training approach with 5 folds. This approach was selected due to the limited amount of training data that was available to the model. The following graph represents the loss function of the training process for each of the classifiers for the five folds. Each fold consists of 20 epochs of training and validation. The best of the five resultant classifier is chosen as the final model.

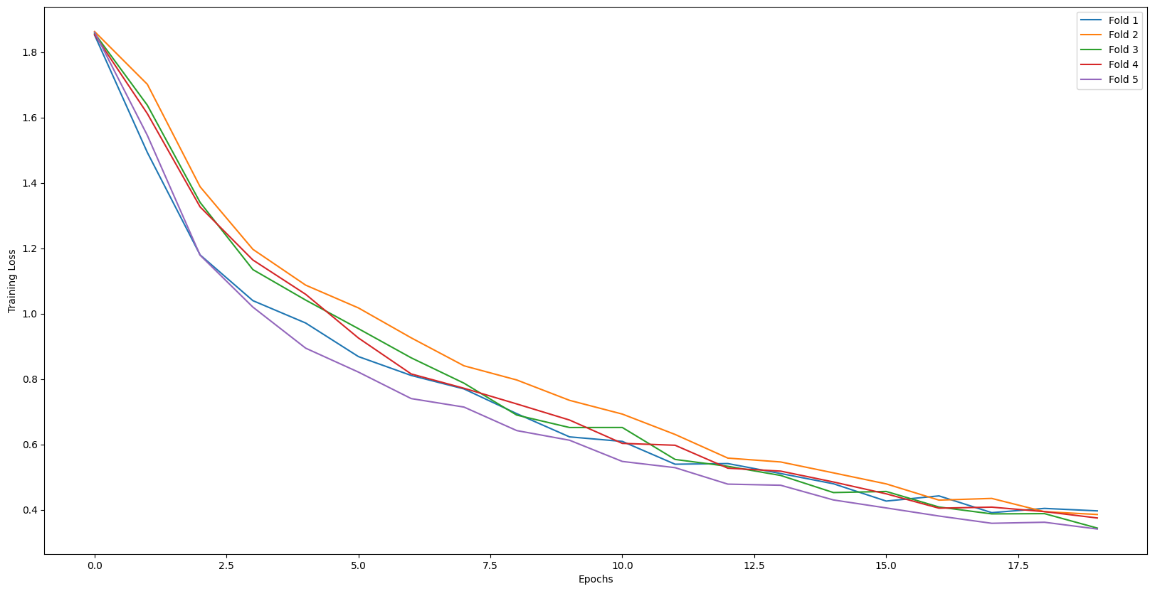

Fig. 5 - Training loss across folds of training using k-fold cross validation

The loss curve for the custom Convolutional Neural Network (CNN) shown below, illustrates the progression of the model's training over epochs. The y-axis represents the loss, a measure of the dissimilarity between the predicted and actual values. A decreasing trend in the curve indicates that the model is learning and optimizing its parameters effectively. The curve provides insights into the convergence and stability of the custom CNN during the training phase.

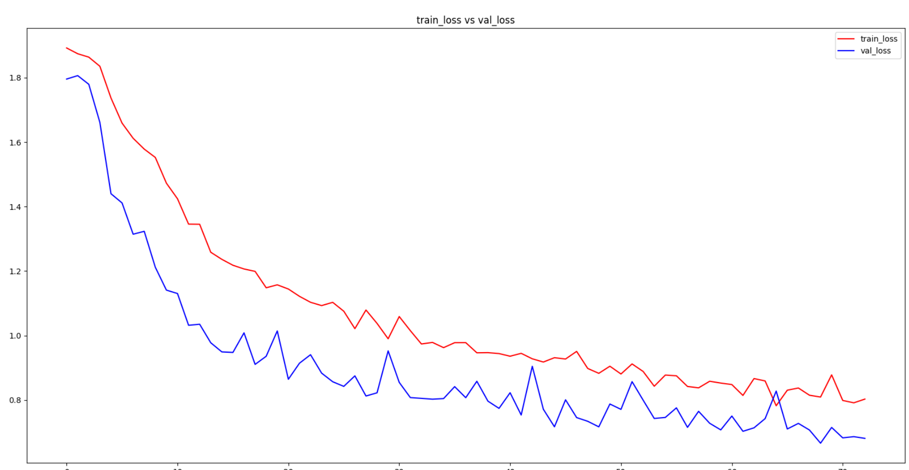

Fig. 6 - Loss Curve for the Custom CNN Model

The accuracy curve for the custom CNN (fig below) showcases the model's ability to correctly classify facial expressions over epochs. The y-axis represents the accuracy, indicating the proportion of correctly classified instances. The rising accuracy curve suggests an improvement in the model's performance, depicting its learning capacity over time. Understanding the accuracy curve is crucial for assessing how well the custom CNN adapts to the complexities of the facial expression recognition task.

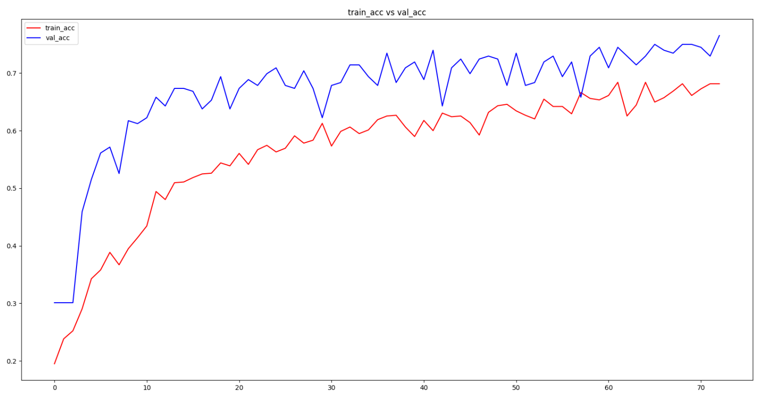

Fig. 7 - Accuracy Curve for the Custom CNN Model

The Receiver Operating Characteristic (ROC) curve for the custom CNN (fig below) provides a visual representation of the model's true positive rate against its false positive rate across varying classification thresholds. The curve closely hugging the top-left corner signifies better performance. The area under the ROC curve (AUC) quantifies the model's overall discriminatory ability, with a higher AUC indicating better performance. Analyzing the ROC curve aids in understanding the trade-offs between sensitivity and specificity in facial expression classification.

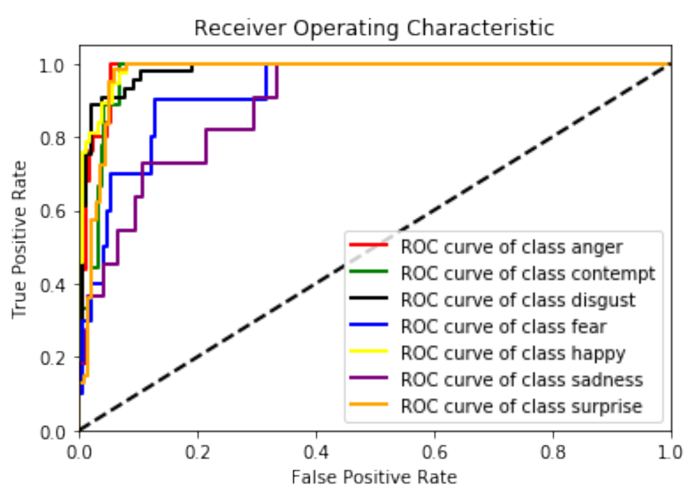

Fig. 8 - ROC Curve for the Custom CNN Model

Similar to the custom CNN, the loss curve for the ResNet model (below fig) depicts the progression of training epochs. A decreasing trend indicates effective learning and optimization. Comparing the loss curves of the custom CNN and ResNet models shows that the ResNet curve is steeper i.e. the loss is decreased with less number of epochs in contrast to our custom CNN

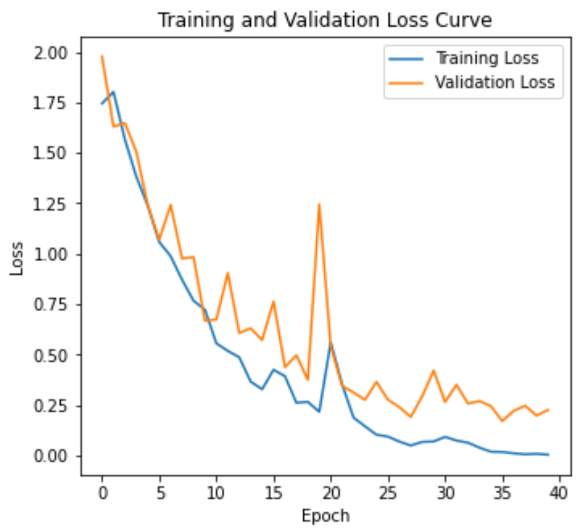

Fig. 9 - Training and Validation Loss Curve for the ResNet Model

The accuracy curve for the ResNet model (fig below) illustrates the model's proficiency in classifying facial expressions over training 40 epochs.

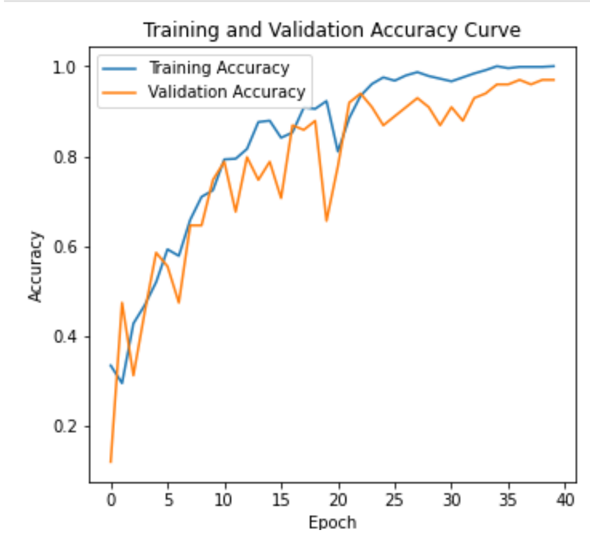

Fig. 10 - Training and Validation Accuracy Curve for the ResNet Model

The classification report presents detailed classification metrics for the ResNet model, including precision, recall, and F1 score for each facial expression class. Precision measures the accuracy of positive predictions, recall quantifies the model's ability to capture all positive instances, and the F1 score balances precision and recall.

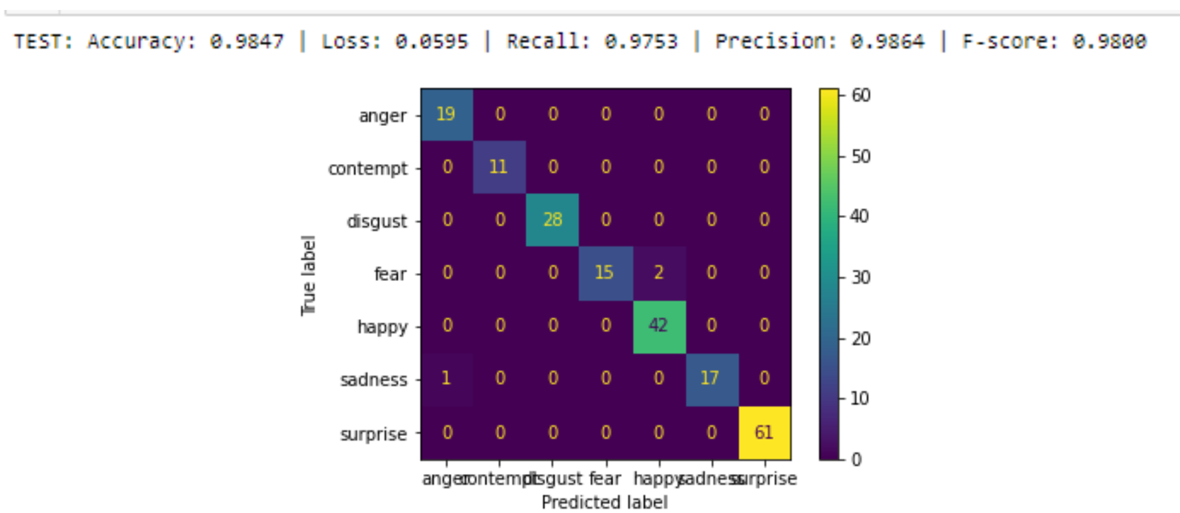

Fig. 11 - Heat map comparison of the predicted and actual classes

We observed that for the custom CNN Model, the Training Accuracy was 92.38%, the Validation Accuracy was 89.48%, and the Testing Accuracy was 90.35%. The below table offers a comprehensive overview of how different preprocessing techniques impact the performance of facial expression recognition models. Comparing accuracy values across models and preprocessing configurations allows for insights into the efficacy of specific preprocessing steps and their influence on overall model performance. The results presented in the table serve as a reference for choosing the most effective combination of preprocessing techniques tailored to the nuances of the CK+ dataset.

S. No.	PRE-PROCESSING TECHNIQUE	TRAINING ACCURACY	VALIDATION ACCURACY	TESTING ACCURACY
1.	NOISE REDUCTION	95.36%	91.74%	90.42%
2.	FACE DETECTION	96.21%	94.89%	92.60%
3.	NORMALIZATION	96.79%	95.48%	93.93%
4.	HISTOGRAM EQUALIZATION	93.57%	90.54%	88.76%
5.	RESNET	99.4%	96.9%	98.4%

Table 1 - Comparison of the Training, Validation and Testing Accuracy for the various Pre-processing Techniques

8. CONCLUSION

The primary objective of this project is to build a Convolutional Neural Network model that can classify facial expressions from images with sufficient accuracy. The various preprocessing techniques used in the work resulted in varying accuracies of their corresponding CNN classifiers. This shows that models learn from a wide range of data points from the images and for different preprocessing techniques the determinant factor might be different. Some preprocessing techniques performed better than others for various reasons. The appropriate preprocessing technique for a dataset might depend on the kind of data that is available to the model. For example, histogram equalization can bring about greater contrast in features in images that are overexposed and grayscale conversion might reveal features that are alike but in different colour channels in the original image.

In comparison with the other preprocessing techniques, normalization performed better in the CK+ dataset. When compared to CNN models, RESNET performed better because of its complex architecture and more sophisticated residual layers. In conclusion, it is observed that the decision to choose one preprocessing technique over another is highly specific to the dataset in hand and a one-size-fits-all solution is unlikely to provide the optimal result in different models. It can also be inferred that preprocessing is a crucial step in the image classification process that can differentiate between a more general classification model and an overfitted model.

9. REFERENCES

1. Andr Teixeira Lopes, Edilson de Aguiar, A. F. De, and Thiago Oliveira-Santos, "Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, Jan. 2017, doi: https://doi.org/10.1016/j.patcog.2016.07.026.

2. C. Pramerdorfer and M. Kampel, "Facial Expression Recognition using Convolutional Neural Networks: State of the Art." Available: https://arxiv.org/pdf/1612.02903.pdf

3. B. Hasani and M. H. Mahoor, "Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields," *arXiv (Cornell University)*, May 2017, doi: https://doi.org/10.1109/fg.2017.99.

4. Jain, Suyog, Changbo Hu, and Jake K. Aggarwal. "Facial expression recognition with temporal modeling of shapes." *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011.

5. T. Kanade, J. Cohn and Y. Tian, "Comprehensive database for facial expression analysis", *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pp. 46-53, 2000. 1, 2.