Emotion Detection from Facial Expressions

Hazem Alsagheer

Chintan Shah

Vasu Patel

***Abstract*—Facial recognition is a vital form of nonverbal communication expressed among humanity. It's used to convey one's emotions which leads to more effective and thoughtful human interactions. Automated facial recognition has significance in various fields including healthcare, and security. This paper provides a deep learning-based approach to classify and label input images by utilizing Convolutional Neural Networks (CNNs). The implemented model captures input images and accurately assigns an emotion label through the extraction of key features executed by the CNNs model. Its promising experimental results and robustness further imply its potential real-world applications. This study underlines the significance of integrating machine learning into enhancing emotion facial detection.**

Keywords— facial recognition, ResNet50, CNN

I. Introduction

Our ability to feel and express emotions defines our core foundations of humanity, enabling us to interact in ways that a machine cannot replicate. Detecting one's emotions has become an indispensable part of our society as it applies in various applications such as health care where robots are trained to detect one's emotions to provide customized therapy to those in need which may include the elderly, and children [1]. It is also used in security applications involving the enhancement of security-based devices such as front door cameras to analyze the behaviour of the detected individual [2]. To apply effective detection, it is essential for machines to be able to understand what shapes our interactions and decision-making processes through effective emotion recognition. However, automated machinery emotion detection usually faces challenges which arise from changes in angles and facial expressions, and lighting conditions. Recent methodologies have addressed such limitations and effectively revolutionized emotion detection. Such methods include Convolutional Neural Networks (CNNs) which represent an advanced image analytical procedure that is set apart from any traditional approach due to its unique automatic extraction of hierarchical features from submitted raw data. Unlike the traditional method, CNN does not require handcrafted feature extraction which allows it to excel in accurate and precise capturing of images over others. This paper utilizes the CNNs model to extract facial features from inputted images through applied emotion detection. The model will apply the appropriate labelling as per the conveyed emotion in the provided image. The designed model aims to address the challenges of traditional methods while maintaining high accuracy and precision in facial recognition and emotion classification.

II. Related Work

Emotion detection is an important part of understanding and interpreting human behaviour, and various machine-learning models have been developed to tackle this challenge. Several academic papers were explored to understand the methodology and current state-of-the-art architectures such as CNN, SVM, and KNN.

In the paper “Face and Facial Expressions Recognition System for Blind People” by J. R. Lee *et al.* [5] builds a system using ResNet50 and CNN architecture to improve facial expression detection for blind people. The methodology included a convoluted neural network (CNN) for unique feature extraction, a pre-trained model using Dlib for facial detection in an image, facial expression recognition performed using ResNet50 and a custom 2D CNN. The authors achieved a training accuracy of 70% and validation accuracy of 60% compared to VGG16 and ResNet50 which only achieved a 34% accuracy. The paper provided a solid foundation for utilizing the Convoluted Neural Architecture using ResNet50.

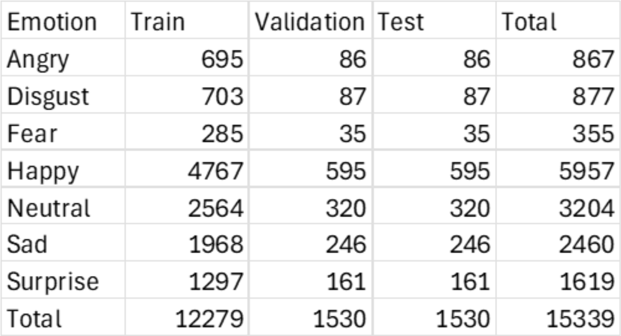
The paper titled “ResEmoteNet: Bridging Accuracy and Loss Reduction in Facial Emotion Recognition” introduces ResEmoteNet which applies a novel approach to implementing Facial Emotion Recognition (FER) through the integration of various neural network components such as Convolutional Neural Networks (CNNs), Squeeze-and-Excitation (SE) Networks, and Residual Networks (ResNets). CNN neural network was utilized to extract facial features through various convolutional layers. Stable learning of the network was achieved by obtaining a normalization batch. Along with ensuring consistent mapping size through average pooling. The model then employs SE to distinguish relevant features from others by emphasizing the important ones and suppressing the rest. ResNet was applied to address vanishing gradient problems and enable learning of deeper networks. The ResEmoteNet model was trained using three well known datasets which are FER2013, RAF-DB, and AffectNet. The model was able to achieve an accuracy of 79.79% on FER2013, 94.76% on RAF-DB, and an accuracy of 72.39% on AffectNet. The model demonstrated superiority over previous related models such as LHC-Net, EmoNeXt, and ResMaskingNet. The overall success of ResEmoteNet was a combination of various methods taken by executing data augmentation through random horizontal flipping and applying optimization with cross-entropy loss with a learning rate scheduler. All of which contributed into a superior novel approach implementing facial recognition [4].

In the paper “Robust real-time emotion detection system using CNN architecture” by S. Jaiswal *et al.* [1] builds a system using CNN architecture to create a real-time facial recognition system. The authors increased the parameters to enable more accurate feature extraction and classification resulting in a 65% accuracy on the training dataset and 74% on the validation dataset. The proposed model performed better than the base CNN models in emotion recognition and robustness as it was validated across multiple dataset.

Overall, CNNs consistently outperform other approaches for facial expression detection in images compared to other methods as proven by the current academic research. Their ability to efficiently extract features, perform accurate detection, and versatility to fine-tune makes them a great foundation for facial expression detection task.

III. Methodology

The goal of this project was to develop a model that is capable of detecting emotions on human faces. To achieve this the RAF-DB (Real-world Affective Face Database) was selected due to it containing 29,672 images with the least noise compared to other similar Facial Emotion Recognition (FER) datasets [5]. Additionally, a subset of the dataset (15,000 images) was available on Kaggle which made prototyping, running and testing the model easy. The dataset contains 7 emotions: Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise and was manipulated to produce a 80%-10%-10% train-test-validation set respectively. The current best model developed is ResEmoteNet which produced a very high test accuracy of 94.76%.



Emotion	Train	Validation	Test	Total
Angry	695	86	86	867
Disgust	703	87	87	877
Fear	285	35	35	355
Happy	4767	595	595	5957
Neutral	2564	320	320	3204
Sad	1968	246	246	2460
Surprise	1297	161	161	1619
Total	12279	1530	1530	15339

Figure 1: Dataset contents

To utilize transfer learning, ResNet50 was chosen as the base model due to its ability to learn complex features and patterns from large datasets which is required for a task like FER [6]. ResNet50 is a 50-layered architecture which utilizes bottleneck residual blocks and skip connections to mitigate the vanishing gradient problem often encountered in deep neural networks [6]. This combined with its pre-training on the ImageNet dataset containing over 14 million images, enables the mode to achieve high accuracy in image classification tasks.

To adapt the model for the RAF-DB dataset, a custom classification head was added which contained a Global Average Pooling (GAP) layer for dimensionality reduction, translation invariance and robustness to overfitting. To conserve computing resources, A fully connected dense layer with 128 units is used along with ReLU activation to introduce non-linearity and allow the model to learn complex features. Through experimentation, a dropout rate of 0.5, L2 regularization of 0.01 yielded the best results and were therefore used to prevent overfitting. A final output layer of 7 units (for 7 classes) with softmax activation is used to determine the probability for each class. The first 40 layers of the model are frozen to preserve the model's pre-trained knowledge of edges, textures and simple shapes. The last 10 layers are trained and fine-tuned to allow the model to learn the features of the RAF-DB dataset and FER in general.

To train the model, two T4 GPUs were used in parallel and mixed-precision training was incorporated to improve training time. Initially, data augmentation such as random horizontal flipping (0.5), random rotation (0.2) and random zoom (0.2) are applied to the training dataset to simulate real-world conditions and improve generalization. The images are also resized to match the input dimensions of ResNet50 which are 224x224 (RGB) pixels. An Adam optimizer with a base learning rate of 0.0001 is used to improve convergence of the model and categorical_crossentropy is chosen as the loss function as it is suitable for multi-class classification. The model was trained for 50 epochs with batch sizes of 128 images.

To monitor the training of the model, a ReduceLROnPlateau is used to dynamically lower the learning rate if validation loss does not improve after 3 consecutive epochs. The difference between validation and training accuracy along with their loss functions are monitored to determine overfitting or convergence. The model with the best validation accuracy is saved using the ModelCheckpoint function.

The hyperparameters mentioned above were fine-tuned by analyzing the validation accuracy of the model. Once an acceptable validation accuracy was acquired, the model was tested for generalizability on the test dataset. Confusion Matrices generated from predictions on the validation and test datasets along with examples of correct and incorrect predictions were used to analyze the strengths and weaknesses of the model.

Furthermore, an application was developed using Python scripts run on Google Collab to allow users to take pictures and detect emotions. Haar Cascade, a pre-trained model on facial recognition is used to identify faces in pictures. If a face is identified it is then resized and fed to the custom-trained ResNet50 model developed in this project. The original image along with the predicted emotion is then outputted to the user along with the confidence of the prediction.

IV. Results and Analysis

The training and validation accuracy graphs (Fig. 2) indicate that the model achieves high training accuracy (90%) but the validation accuracy plateaus at 81%. This suggests that the model has learnt well but has overfitted which is also indicated by a high validation loss of 0.715. The accuracy and loss on the testing dataset were 78.3% and 0.8 respectively, also indicating that the model is generalizing decently to unseen data but still has room for improvement.

The model demonstrates strong performance for the "Happy" class in both datasets, with high precision, recall and F1 scores while classes like "Fear" and "Disgust" have significantly lower metrics. The confusion matrices (Fig. 3

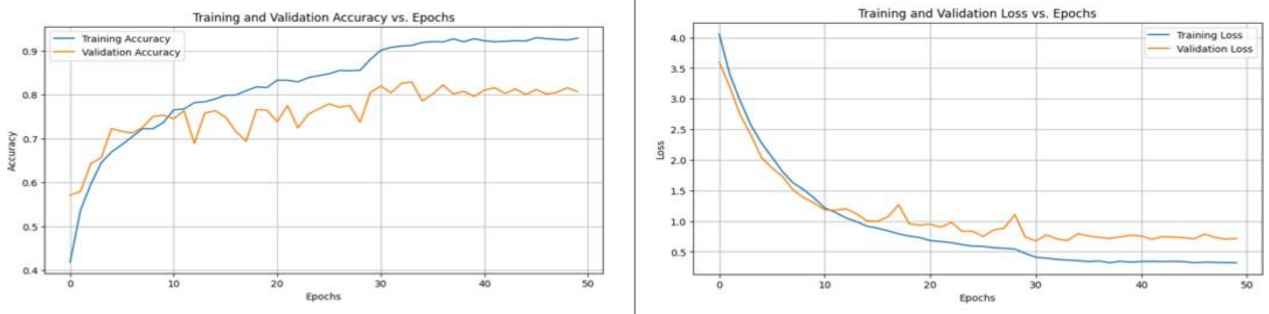
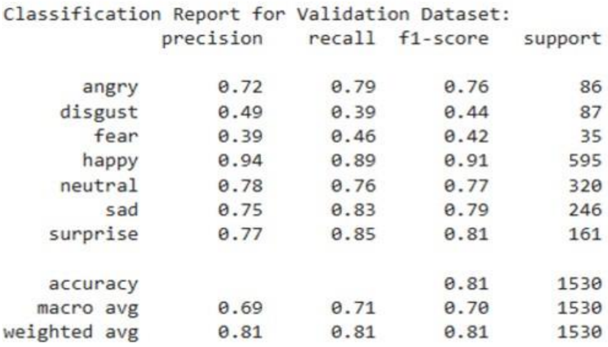


Fig. 2: Training and validation Accuracy Vs Epoch (left) and Loss Vs Epoch (right)



```
Classification Report for Validation Dataset:
              precision    recall  f1-score   support

       angry       0.72      0.79      0.76        86
     disgust       0.49      0.39      0.44        87
        fear       0.39      0.46      0.42        35
       happy       0.94      0.89      0.91       595
     neutral       0.78      0.76      0.77       320
         sad       0.75      0.83      0.79       246
    surprise       0.77      0.85      0.81       161

    accuracy                           0.81      1530
   macro avg       0.69      0.71      0.70      1530
weighted avg       0.81      0.81      0.81      1530
```

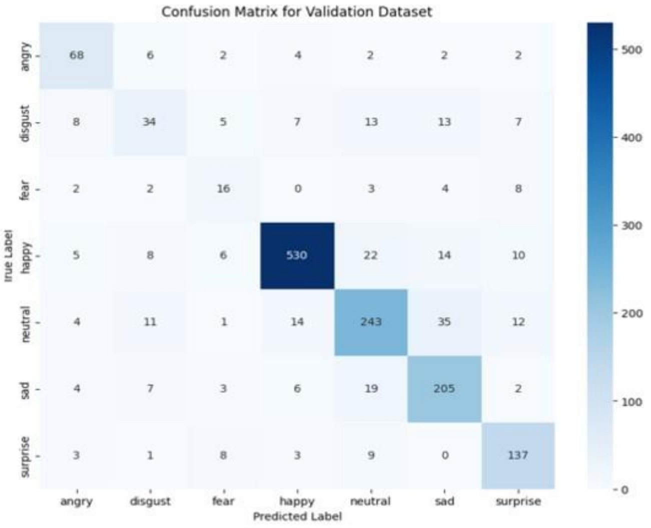
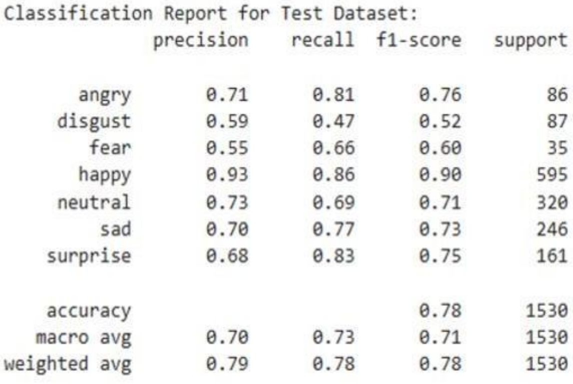


Fig. 3: Classification Report (left) and Confusion Matrix (right) for validation dataset



```
Classification Report for Test Dataset:
              precision    recall  f1-score   support

       angry       0.71      0.81      0.76        86
     disgust       0.59      0.47      0.52        87
        fear       0.55      0.66      0.60        35
       happy       0.93      0.86      0.90       595
     neutral       0.73      0.69      0.71       320
         sad       0.70      0.77      0.73       246
    surprise       0.68      0.83      0.75       161

    accuracy                           0.78      1530
   macro avg       0.70      0.73      0.71      1530
weighted avg       0.79      0.78      0.78      1530
```

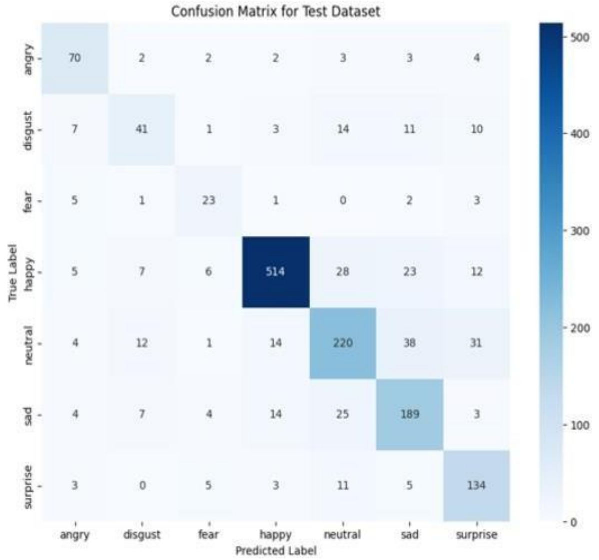


Fig. 4: Classification Report (left) and Confusion Matrix (right) for test dataset

and 4) indicate that there are frequent misclassifications between similar datasets. In particular, “Neutral” classes are often misclassified as “Happy” or “Sad” and “Sad” classes are often misclassified as “Happy” or “Neutral”. “Surprise” performs well on the validation dataset but is often misclassified on the test dataset indicating a need for better feature representation or balanced data distribution.

By analyzing the images in Fig. 5 and 6, it can be seen that the model has high confidence on emotions like “Happy” and “Neutral” which have easily identifiable features such as wide smiles or relaxed expressions. However, it struggles with subtle or overlapping features in emotions such as “Angry”, “Disgust” and “Fear” and often misclassify them due to shared characteristics like clenched jaws or wide open mouths. Inadequate images due to low resolution, occlusion or partial faces also confuse the model leading to misclassification. These results suggest that the model is overly reliant on prominent facial features and has difficulties handling subtle variations.

V. Conclusion

In conclusion, the project successfully developed a ResNet50-based model which trained on the RAF-DB dataset, achieving reliable accuracy in recognizing emotions from facial images. The model demonstrates strong performance for emotions with distinct facial features such as “Happy” or “Neutral”, but struggles with subtle features in emotions such as “Fear”, “Disgust” and “Angry”. Despite achieving decent generalization with a validation accuracy of 81% and test accuracy of 78%, the model is still affected by overfitting and misclassification.

To address the issue of class imbalance and improve feature extraction capabilities, the model should be trained on the full RAF-DB dataset which includes 29,672 images.

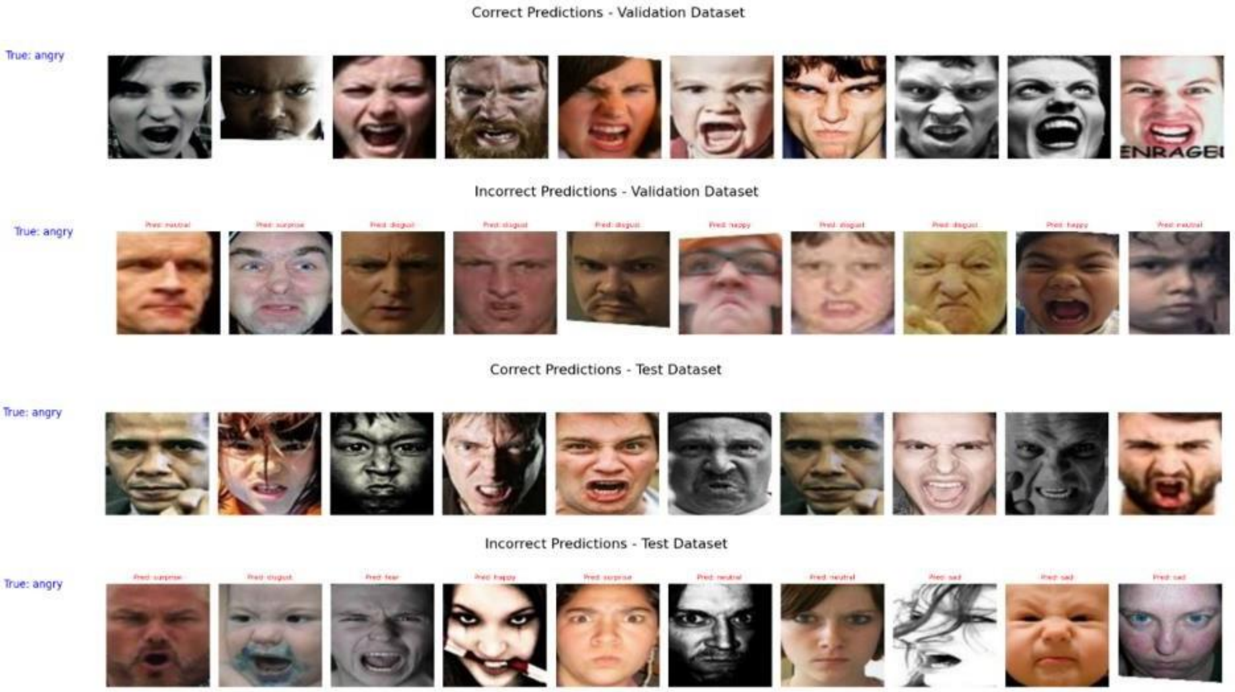


Fig. 5: Correct and incorrect prediction examples on validation and test datasets. Full grid for all emotions provided in attached CPS843_Custom_RAF_DB.ipynb file.

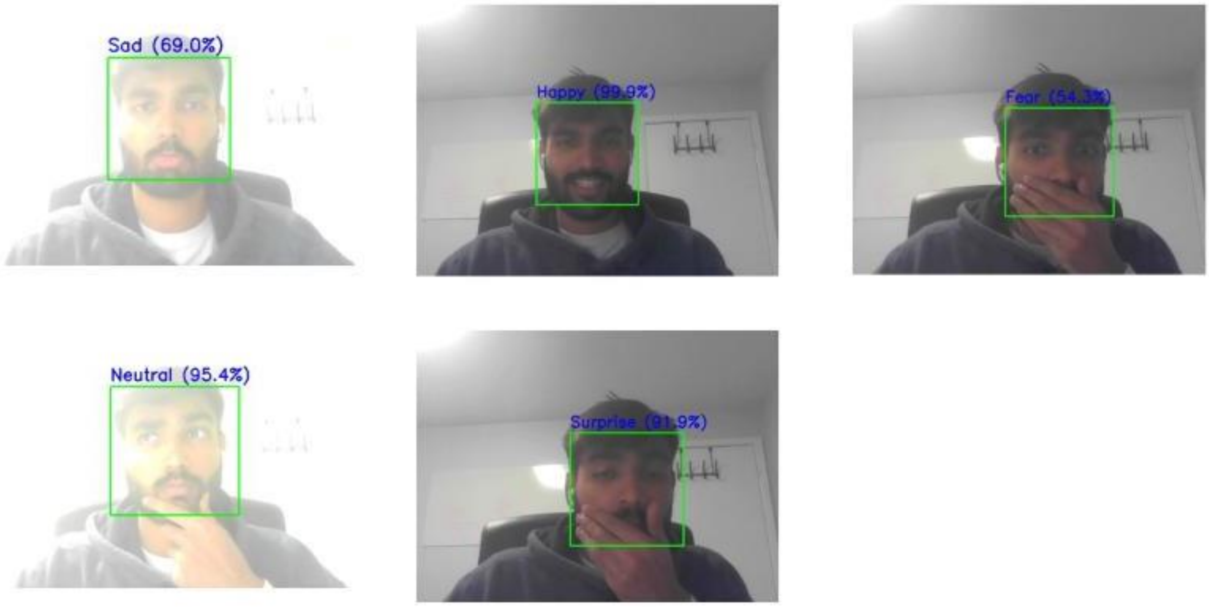


Fig. 6: Predictions from the Emotion Detector application.

Additionally, Vision Transformers (ViTs), dual-directional attention networks and locally forced multi-headed attention mechanisms can be explored. These approaches would better capture spatial relationships and subtle details in facial recognition tasks. To improve the generalization and robustness of the model further data augmentation techniques can be utilized. By implementing these strategies, the model can get a closer test accuracy to the current state of the art and become more effective in real-world applications.

REFERENCES

[1] S. Jaiswal and G. C. Nandi, "Robust real-time emotion detection system using CNN architecture," *Neural Computing and Applications*, vol. 32, no. 15, pp. 11253–11262, Oct. 2019. doi:10.1007/s00521-019-04564-4

[2] A. Bhattacharya, P. Dash, M. Jain, and A. Jothimani, "Smart Home Security System using Emotion Detection," *International Research Journal of Engineering and Technology*, vol. 7, no. 5, pp. 2179 - 2184, May 2020.

[3] J. R. Lee, K.-W. Ng, and Y.-J. Yoong, "Face and Facial Expressions Recognition System for Blind People Using ResNet50 Architecture and CNN", JIWE, vol. 2, no. 2, pp. 284–298, Sep. 2023.

[4] A. K. Roy, H. K. Kathania, A. Sharma, A. Dey, and Md. S. A. Ansari, "ResEmoteNet: Bridging Accuracy and Loss Reduction in Facial Emotion Recognition," arXiv, 2024. doi: 10.48550/arxiv.2409.10545

[5] O. S. Ekundayo and S. Viriri, "Facial Expression Recognition: A Review of Trends and Techniques," *IEEE Access*, vol. 9, pp. 136944–136973, 2021, doi: 10.1109/ACCESS.2021.3113464

[6] S. Vats, A. N. Singh, V. Kukreja, and R. Sharma, "Leveraging Pre-trained Deep Learning Models for Orange Leaf Disease Classification," in 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), IEEE, 2024, pp. 1–4. doi: 10.1109/I2CT61223.2024.1054406.