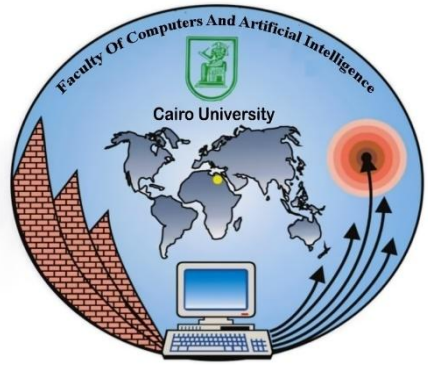

Faculty of Computers and Artificial Intelligence Cairo University

Optimizing Facial Expression Recognition Features with Pre-trained CNNs

Implemented By:

Mahmoud Wael

1. Abstract

This paper aims to enhance the efficiency of facial expression recognition by optimizing features extracted from the FER-2013 dataset. Leveraging the pre-trained MobileNetV2 model from ImageNet, the Ant Colony Optimization algorithm is applied to perform feature selection on the FER-2013 dataset. Additionally, the Synthetic Minority Over-sampling Technique (SMOTE) is employed to address the class imbalance issue in the dataset, improving the representativeness of the training data.

The selected features are then passed as an input to a Random Forest classifier for accurate expression classification.

Despite the utilization of advanced techniques, achieving high accuracy in facial emotion expression classification remains challenging. Further exploration and refinement of methodologies are necessary for significant improvements.

Keywords: MobileNetV2, Ant Colony, Feature Selection, Feature Extraction.

2. Introduction

The rapid advancement of human-computer interaction and pattern recognition, coupled with frequent updates in computer hardware, has empowered people to delegate complex tasks to computers, meeting various life and market demands. This has brought significant convenience to society. A recent development in this field is facial expression recognition, an intelligent method for human-computer interaction. It has a wide range of applications, including VR games, medical care, online education, driving, security, and more. For instance, some cameras now feature a smile mode that automatically takes a photo when a smile is detected, enhancing the user experience. In certain European countries, facial expression recognition is used to monitor the emotional fluctuations of elementary school students in classrooms, aiding in individualized treatment and analysis of their learning progress. High-end car brands like Toyota's Lexus employ facial expression recognition to monitor driver's eyes and expressions to prevent fatigue related accidents. Facial expressions are a vital means of conveying emotions, often revealing an individual's inner thoughts. The primary purpose of facial expressions is to capture the subject's emotional changes through their facial emotions. Compared to other communication methods, facial expressions offer a wider range of diversity and can inadvertently reveal one's genuine feelings. In 1971, Ekman first categorized expressions into six basic forms: sadness, happiness, fear, disgust, surprise, and anger, with a normal expression later added to the FER-2013 dataset. These expressions can be challenging to manually categorize, and humans can classify faces with an accuracy of 63% ±5% among the seven emotions. Facial expression recognition (FER) plays a pivotal role in comprehending human emotions and sentiments. politics, and medical domains. Automated FER systems have been developed and employed to discern human emotions, but they've encountered significant challenges within the realm of machine learning, primarily stemming from the substantial variations within the same class. The initial models utilized conventional techniques like Support Vector Machines (SVM), Bayes classifiers, Fuzzy Techniques, Feature Selection, and Artificial Neural Networks (ANN). However, these models still grappled with limitations that critically affected accuracy, including subjectivity, occlusion, pose variations, low resolution, scale differences, illumination fluctuations, and more. The introduction of Convolutional Neural Networks (CNN) has significantly elevated the accuracy of FER systems. In recent years, deep learning algorithms have emerged as the most effective approach to achieving top notch FER results. Various datasets, such as FER2013, CK+, JAFFE, and FERG, have been utilized to train, test, and validate FER models. To enhance model accuracy, researchers have sometimes combined these datasets, although each dataset comes with its own limitations and challenges, influencing the performance of the models trained on them. Facial expression recognition methods today fall into two primary categories: traditional manual approaches and network models employing deep learning.

While the traditional method has seen widespread use, it is notably constrained in practical applications. On the other hand, leveraging deep learning for facial expression classification

typically involves harnessing the power of strong supervision methods to capture emotional features from extensive sample data. Convolutional Neural Networks (CNNs) are a powerful tool for facial expression recognition, but they can encounter several challenges when trying to determine facial features and a person's current emotional state:

1. Variability in Facial Expressions: People express emotions differently, and even within the same emotional category (e.g., "happiness"), there can be significant variations in facial expressions. CNNs may struggle to capture the full range of these expressions accurately.
2. Occlusion: When parts of the face, such as the eyes, mouth, or other key regions, are partially or fully obscured, it can hinder CNN's ability to identify facial features essential for recognizing emotions.
3. Illumination Changes: Changes in lighting conditions, such as shadows or strong illumination, can affect the appearance of facial features and the interpretation of emotions.
4. Subjectivity and Ambiguity: The interpretation of facial expressions can be subjective and context dependent. What may be perceived as one emotion by one person could be interpreted differently by another, making the task challenging.
5. Data Imbalance: Imbalanced datasets with more examples of some emotions than others can lead to biased models that are better at recognizing overrepresented emotions while performing poorly on underrepresented ones.
6. Cross-Cultural Variation: Different cultures may express emotions in distinct ways, and CNNs trained on one cultural group's data may not generalize well to others.
7. Real-time Processing: In real-time applications, such as emotion recognition from live video streams, CNNs need to process data quickly and efficiently, which can be computationally demanding. To address these challenges, researchers continue to develop and refine CNN models for facial expression recognition, and they often use large, diverse datasets and apply techniques like data augmentation and fine-tuning to improve performance. Additionally, combining CNNs with other techniques can help enhance emotion recognition accuracy.

In our study, we encountered significant class imbalance in the FER-2013 dataset, which could bias the model towards overrepresented classes. To address this challenge, we employed the SMOTE (Synthetic Minority Over-sampling Technique), a popular method for generating synthetic samples in underrepresented classes.

3. Related Work

Facial Expression Recognition (FER) has been extensively researched and numerous different approaches have been suggested. This exploration has led to various innovative methods and techniques due to the importance of the FER as a computer vision task and it's benefits to humanity. Those who choose to work on the FER often strive to enhance the existing performance in their own way, even by fine-tuning an existing model or changing the architecture of a pre-trained model or augmenting the used datasets or even inventing a new model from scratch and so on. There are several contributions that have been made in the FER task in different ways as we mentioned, and we are going to discuss some of these contributions in this section.

1. One of the problems in FER is that training neural networks is significantly more difficult as most of the existing databases have a small number of images or video sequences for certain emotions, and most of these databases contain still images that are unrelated to each other which makes the task of sequential image labeling more difficult. While Inception and ResNet have shown remarkable results in FER, these methods don't extract the temporal relations of the input data, so to capture the spatial intricacies of facial images and the temporal dependencies across video frames, Hasani introduced a method which extracts temporal relations of consecutive frames in a video sequence using 3D convolutional networks and Long Short-Term Memory (LSTM). They extract and incorporate facial landmarks in their proposed method that emphasizes more expressive facial components which improve the recognition of subtle changes in the facial expressions in a sequence. This method was evaluated using four facial expression datasets (CK+, MMI, FERA, and DISFA) for expressions classifications, including cross-database classification tasks. [1] He also proposed a CNN architecture using a function with bounded derivative instead of a simple shortcut path in the residual units for automatic recognition of facial expressions. This method proposed adaptive complex mapping results in a shallower network with less numbers of training parameters compared to ResNet. It was evaluated on the AffectNet, FER2013, and Affect-in-Wild datasets. [2]

2. Georgescu proposed a method that combines automatically extracted features from a Convolutional Neural Network (CNN) with handcrafted features obtained using the bag-of-visual-words (BOVW) approach. This combination forms a model for Facial Emotion Recognition (FER). To generate the automatic features, the approach involves a two-step process. Once these two types of features are fused, a local learning framework is applied to predict the class label for each test image. This local learning framework consists of three steps. First, a k-nearest neighbors model is used to select the nearest training samples for a given test image. Second, a one-versus-all support vector machines (SVM) classifier is trained based on the selected training samples. Finally, the SVM classifier is utilized to predict the class label for the specific test image it was trained for. he results of experiments conducted on three datasets, namely the 2013 FER Challenge

dataset, the FER+ dataset, and the AffectNet dataset, indicate that this approach achieves state-of-the-art performance. Specifically, it achieves top accuracy rates of 75.42% on the FER 2013 dataset, 87.76% on the FER+ dataset, 59.58% on the AffectNet eight-way classification, and 63.31% on the AffectNet seven-way classification. [3]

3. In contrast to previous methods that focus on recognizing human emotions primarily through facial expressions, speech, or gestures, some researchers have recognized the untapped potential of incorporating contextual information from the surrounding environment. Hoang introduced the concept that background data, in the broader sense, can serve as supplementary cues for emotion recognition. He introduced a methodology that leverages the visual connections between the primary subject and surrounding objects in the scene to infer facial emotions. This approach combines both spatial and semantic characteristics of objects in the scene, evaluating the impact of all context-related elements along with their respective properties (positive, negative, or neutral) on the primary subject through a customized attention mechanism. The model integrates these derived features with the overall scene context and the physical characteristics of the subject to make predictions about their emotional states. The experimental results demonstrate that this approach achieves exceptional performance on the CAER-S dataset. [4]

4. Recognition of human emotion recognition using audio and visual features is also studied in some previously proposed work. Schoneveld used deep feature representations of the audio and visual modalities to improve the accuracy of the FER task by fusing the feature representation of the visual and audio modalities based on a model-level fusion strategy and a recurrent neural network is then used to capture the temporal dynamics. [5]

5. The majority of existing methods primarily concentrate on the multi-modal feature learning and fusion strategy, which pay more attention to the characteristics of a single video and ignore the correlation among the videos. To explore this correlation, Zhoa took an innovative path by examining audio features using speech-spectrogram and Log Mel-spectrogram and evaluated facial features with different CNNs and different emotion pretrained strategies. [6]

6. Szegedy introduced GoogLeNet a 22 layers deep network which consists of multiple "inception" layers inception applies several convolutions on the feature map in different scales. [7]

7. Octavio Arriaga, Matias Valdenegro-Toro and Paul Plöger build a framework (or a system) for designing real-time CNNs.The models were validated by creating a real-time vision system which accomplishes the tasks of face detection, gender classification and emotion classification simultaneously in one blended step using their proposed CNN architecture. The reported accuracies were 96% in the IMDB gender dataset and 66% in the FER-2013 emotion dataset. [8]

8. Ali Ghofrani, Rahil Mahdian Toroghi and Shirin Ghanbari apply facial expression recognition which contains of two different stages: 1. Face detection, 2. Emotion Recognition. For the first stage, an MTCNN (Multi-Task Convolutional Neural Network)

has been employed to accurately detect the boundaries of the face, with minimum residual margins. The second stage leverages a ShuffleNet V2 architecture which can tradeoff between the accuracy and the speed of model running, based on the users' conditions. The dataset used was the FER 2013 on Kaggle. [9]

9. Tee Connie, Mundher Al-Shabi, Wooi Ping Cheah and Michael Goh apply a facial expression recognition task using CNN. They focus on achieving good accuracy while requiring only a small sample data for training. Scale Invariant Feature Transform (SIFT) features are used to increase the performance on small data as SIFT doesn't require extensive training data to generate useful features. The proposed approach is tested on the FER-2013 and CK+ datasets. Results demonstrate the superiority of CNN with Dense SIFT over conventional CNN and CNN with SIFT. The accuracy even increased when all the models are aggregated which generates state-of-art results on FER-2013 and CK+ datasets, where it achieved 73.4% on FER-2013 and 99.1% on CK+. [10]

10. Minchul Shin, Munsang Kim and Dong-Soo Kwon presented a baseline (CNN) structure and image preprocessing methodology to improve facial expression recognition algorithm using CNN. To analyze the most efficient network structure, they investigated four network structures that are known to show good performance in facial expression recognition. They also investigated the effect of input image preprocessing methods, trained 20 different CNN models (4 networks × 5 data input types) and verified the performance of each network with test images from five different datasets. The experiment result showed that a three-layer structure consisting of a simple convolutional and a max pooling layer with histogram equalization image input was the most efficient. They used variant datasets like SFEW2.0, SFEW and FER2013 datasets. [11]

11. A. Nasuha, F. Arifin, A. S. Priambodo, N. Setiawan and N. Ahwan applied emotion classifier based on facial features. Here, they used CNN to extract facial features from input images and classify them into 7 classes. Convolution is applied, instead of the ordinary convolution in CNN, to reduce the number of trainable parameters so that the overall architecture of CNN can be made as simple as possible without compromising the accuracy. work in real time. This method achieves an accuracy of 66% on 3.589 input images using FER 2013 dataset. [12]

12. Chengkun Shi, Xiaoqing Zhou, ZhiCheng Zhang, Jie Xu invented a new way to apply facial expression recognition method that utilizes the 2D DCT, k-means algorithm and vector matching. This technique is based on two main ideas: (i) complicated facial expression categories such as "anger" and "sadness", may be divided into

several subcategories with different sub feature spaces where the recognition task can be performed with higher accuracy and (ii) the k-means algorithm may be used to cluster these subcategories. [13]

4. The Dataset [14]

The FER-2013 dataset, which was unveiled at the International Conference on Machine Learning (ICML) in 2013 by Pierre-Luc Carrier and Aaron Courvill is a dataset provided by Kaggle. In this dataset, each face has been categorized based on emotion categories, where the FER-2013 dataset is a grayscale image measuring 48 pixels by 48 pixels for each image. The total FER-2013 dataset is 35,887 consisting of 7 different types of micro expression and marked with labels based on 7 different classifications starting from the index label 0 to 6.

A micro expression is a facial expression that can easily observe and distinguish it as a communication method in social psychology These facial expressions serve as conveyors of emotional information, revealing our intentions and objectives, and playing a pivotal role in human interactions. The ability to recognize and understand facial expressions automatically facilitates the intended communication. The process in the classification of human facial expressions consists of three stages: face detection, feature extraction, and facial expression classification. The dataset consists of 7 basic human expressions:

1. Happiness:

A smiling expression is a facial expression that often signifies feelings of joy or liking something. It is characterized by the lifting of the cheek muscles and the corners of the lips, forming a pleasant and cheerful smile.

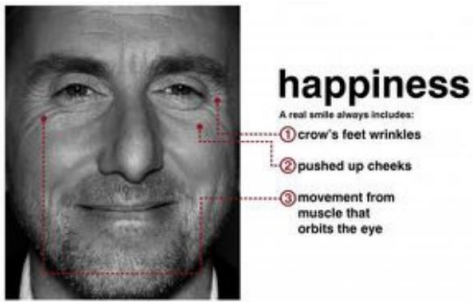

Figure 1: Facial Expression of Happiness [15]

2. Anger:

Facial expressions of anger result from a disappointment between one's expectations and the actual reality they encounter.
This emotional expression is typically manifested by a distinct pattern: the inner eyebrows converge and tilt downward, the lips narrow, and the eyes assume a sharp, focused appearance when gazing.

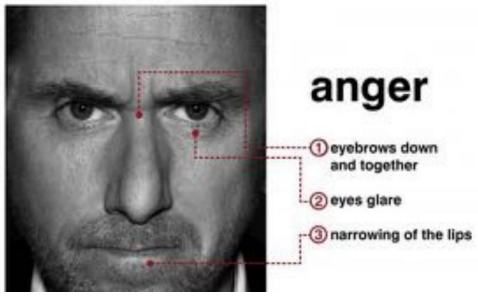

Figure 2: Facial Expression of Anger [15]

3. Sadness:

A face that shows sadness is typically associated with feelings of letdown, disappointment, or a sense of lacking something. This emotion is often identifiable through certain characteristics, such as a lack of focus in the eyes, a downward pull of the lips, and a drooping upper eyelid.

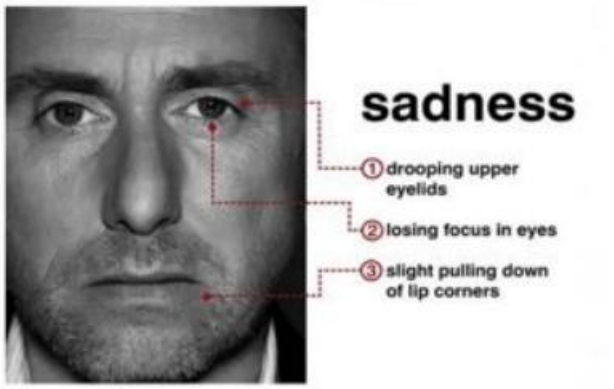

Figure 3: Facial Expression of Sadness [15]

4. Fear:

When an individual encounters a situation they find challenging or experiences fear in a frightening environment, they display an expression of fear. This fearful expression is characterized by the simultaneous raising of both eyebrows, tightening of the eyelids, and horizontal opening of the lips on the person's face.

Figure 4: Facial Expression of Fear [15]

5. Disgust:

A person who expresses his face in a state of disgust due to seeing something not common or listening to information that is not worth hearing. An expression of disgust will be recognized when a person's face in the area of the nose bridge is wrinkled and the upper lip rises.

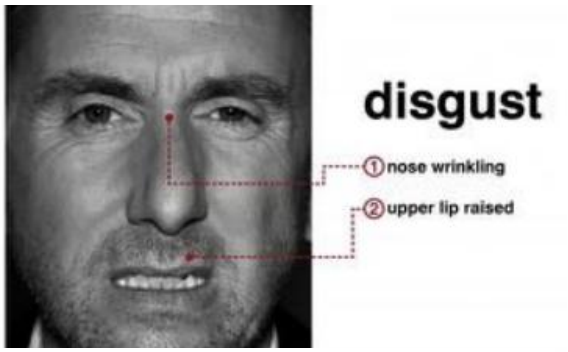

Figure 5: Facial Expression of Disgust [15]

6. Surprise:

Surprise expressions occur when an individual is confronted with an unforeseen, sudden, or significant event or message for which they had no prior knowledge. These expressions are typically characterized by a startled facial appearance, including raised eyebrows, wide-open eyes, and a reflexive opening of the mouth.

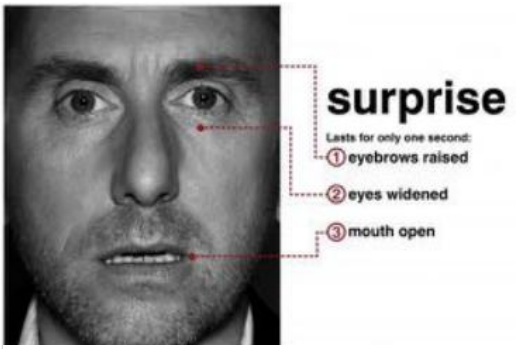

Figure 6: Facial Expression of Surprise [15]

7. Contempt:

Contempt is the facial expression typically associated with an individual displaying arrogance and a lack of respect towards others. It often involves underestimating or belittling others. This emotion is conveyed through a subtle movement that elevates one corner of the lips.

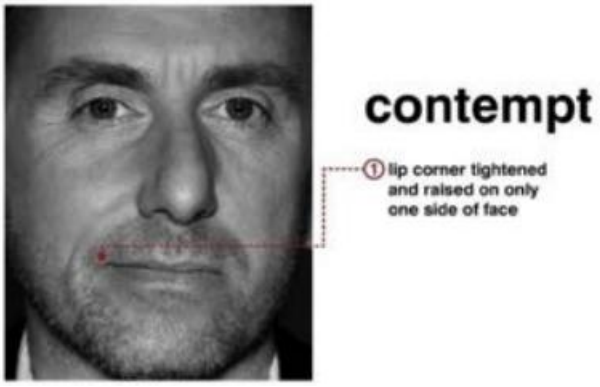

Figure 7: Facial Expression of Contempt [15]

5. Proposed Model

A. Model Architecture:

The proposed model leverages the MobileNetV2 architecture, pretrained on the ImageNet dataset. MobileNetV2 is renowned for it's efficiency and lightweight design, making it well-suited for facial expression recognition tasks. MobileNetV2 primarily consists of repeated inverted residual blocks, each comprised of a depthwise separable convolution with a 3x3 kernel, batch normalization and ReLU activation and a linear bottleneck layer (1x1 convolution).

B. Training Strategy:

To capitalize on the knowledge encoded in the pretrained MobileNetV2 model, the trainable parameters are frozen during training. This strategic decision enables the model to leverage the learned weights, providing a foundation for more accurate facial expression recognition.

C. Dataset Utilization:

The FER-2013 dataset undergoes a preprocessing step involving the normalization of pixel values. This prepares the data for optimal interaction with the MobileNetV2 model. The normalized data is then seamlessly fed into the network for subsequent feature extraction and classification.

D. Feature Extraction:

The core of the proposed model lies in it's ability to extract the features from the images through convolutional neural network (CNN) architecture. The pretrained MobileNetV2 serves as a feature extractor, capturing intricate patterns and expressions inherent in facial images.

E. Hyperparameters:

The model inherits the hyperparameters from the pretrained MobileNetV2, including learning rates, batch sizes, and other configuration settings. This approach ensures consistency with the original architecture and facilitates seamless integration into the facial expression recognition pipeline.

F. Addressing Class Imbalance:

To tackle the class imbalance issue in the FER-2013 dataset, we incorporated SMOTE after extracting the features by the MobileNetV2 model. SMOTE generates synthetic samples by interpolating between existing samples in the minority class, effectively balancing the dataset. This preprocessing step improved the diversity and representativeness of the data, enabling the model to learn features from all classes more effectively.

G. Optimization Algorithms:

Feature selection, a critical aspect of optimizing facial expression recognition, is achieved through the implementation of the Ant Colony Optimization algorithm. This algorithm efficiently navigates the feature space, selecting the most influential features from those extracted by the MobileNetV2 model. This strategic feature selection contributes to the model's overall efficiency and accuracy.

H. Validation:

To assess the robustness of the proposed model, validation is performed on the test dataset. The model's performance, along with the effectiveness of the Ant Colony Optimization algorithm in feature selection, is rigorously evaluated. This comprehensive validation process ensures the model's capability to generalize to unseen data and effectively capture the nuances of facial expressions.

6. Experimental Results

The experiments were conducted on a P100 GPU. The feature extraction process involved flattening the output of the MobileNetV2 model, followed by passing it through three dense layers.

The Ant Colony Optimization algorithm parameters were set as follows:

- Number of Ants = 20
- Alpha = 1.1
- Beta = 1.5
- The maximum number of iterations = 3
- The Exploration Factor (Q0) = 0.9
- The pheromone decay rate (rho) = 0.2

A. Dataset:

The FER-2013 dataset consists of grayscale images, each of size 48×48 pixels, with a total of 35,887 images. It consists of seven distinct micro-expression categories, labeled from 0 to 6. Preprocessing involved normalizing the pixel values.

B. CNN Architecture:

The MobileNetV2 architecture was employed for feature extraction. It utilizes inverted residual blocks, a design that helps maintain computational efficiency while capturing essential information and linear bottlenecks which enhance it's ability to learn features effectively. This model was initially trained on the ImageNet dataset, showcasing its proficiency in recognizing a wide range of visual patterns. The input of the architecture takes images of size [48×48×3]. The MobileNetV2 model, with its top layers excluded, was paired with a flattening layer applied to its output to prepare the features for further processing.

C. Feature Extraction:

Features were extracted from both training and testing images using the MobileNetV2 model's predict function. To address class imbalance, SMOTE was applied to resample the extracted training features and labels, ensuring a more balanced dataset for training. The model was modified by adding a flattening layer to the output of the MobileNetV2 model. This layer converts the multi-dimensional feature maps into a one-dimensional vector, making them suitable for input into subsequent processing steps such as classification or optimization. The optimization algorithm then selects the most important features from the extracted features.

D. The Classifier Configuration:

A Random Forest classifier with 90 estimators was employed. The optimization algorithm determined the optimal number of features during training on the training data. The classifier was then fitted on these features and evaluated on the test data, resulting in an accuracy measurement.

E. Results:

The model achieved an accuracy 35%, a Precision of 0.381, a recall of 0.350 and F1 score of 0.326 indicating the difficulty of accurately classifying facial expressions. After incorporating SMOTE, we observed a notable improvement in the model's performance. The overall accuracy increased to 49.7%, with significant gains in the precision, recall and F1 score of 0.498, 0.498 and 0.493 respectively. This highlights the importance of addressing class imbalance in FER tasks to achieve fair and robust results

7. Conclusion

In this study, we explored the complex field of recognizing facial expressions, leveraging the power of deep neural networks for feature extraction and learning. While these networks offer unparalleled effectiveness, their extensive training times, especially with modest hardware, pose a significant challenge.

Our focus was on optimizing facial expression recognition features using a pre-trained MobileNetV2 model and the ant colony optimization algorithm. Employing the convolutional mastery of MobileNetV2 and coupling it with ant colony optimization, our objective was to extract and select the most crucial features, we aimed to enhance the classification accuracy of our system through experimenting our method on the FER-2013 dataset which consisting of 35,887 grayscale images across seven emotion classes.

Incorporating SMOTE to balance the classes proved essential for improving classification fairness and accuracy. The results demonstrate that addressing data imbalance is a critical step in enhancing the performance of facial expression recognition systems.

In conclusion, our exploration into facial expression recognition highlights the significance of careful parameter tuning and feature optimization. While challenges persist, the promise of deep learning remains steadfast, offering a practical and efficient avenue for future advancements in emotion recognition technology.

8. References

[1] B. Hasani and M. H. Mahoor, ‘‘Facial expression recognition using enhanced deep 3D convolutional neural networks’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 30 - 40. [Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks | IEEE Conference Publication | IEEE Xplore

[2] B. Hasani, P. S. Negi, and M. Mahoor, ‘‘BReG-NeXt: Facial affect computing using adaptive residual networks with bounded gradient’’ IEEE Trans. Affect. Comput., early access, Apr. 13, 2020. BReG-NeXt: Facial Affect Computing Using Adaptive Residual Networks With Bounded Gradient | IEEE Journals & Magazine | IEEE Xplore

[3] M. Georgescu, R. T. Ionescu, and M. Popescu, ‘Local learning with deep and handcrafted features for facial expression recognition’’ IEEE Access, vol. 7, pp. 64827 - 64836, 2019. Local Learning With Deep and Handcrafted Features for Facial Expression Recognition | IEEE Journals & Magazine | IEEE Xplore

[4] M.-H. Hoang, S.-H. Kim, H.-J. Yang, and G.-S. Lee, ‘‘Context-aware emotion recognition based on visual relationship detection’’ IEEE Access, vol. 9, pp. 90465 - 90474, 2021. Context-Aware Emotion Recognition Based on Visual Relationship Detection | IEEE Journals & Magazine | IEEE Xplore

[5] L. Schoneveld, A. Othmani, and H. Abdelkawy, ‘‘Leveraging recent advances in deep learning for audio-visual emotion recognition’’ pp. 1 - 7, Jun. 2021. [2103.09154] Leveraging Recent Advances in Deep Learning for Audio-Visual Emotion Recognition (arxiv.org)

[6] H. Zhou, D. Meng, Y. Zhang, X. Peng, J. Du, K. Wang, and Y. Qiao, ‘‘Exploring emotion features and fusion strategies for audio-video emotion recognition’’ in Proc. Int. Conf. Multimodal Interact., 2019, pp. 562 - 566. C-GCN: Correlation Based Graph Convolutional Network for Audio-Video Emotion Recognition | IEEE Journals & Magazine | IEEE Xplore

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. ‘‘Going deeper with convolutions’’. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1 - 9, 2015. [1409.4842] Going Deeper with Convolutions (arxiv.org)

[8] Octavio Arriaga, Matias Valdenegro-Toro and Paul Plöger” Real-time Convolutional Neural Networks for Emotion and Gender Classification” [Submitted on 20 Oct 2017]. [1710.07557] Real-time Convolutional Neural Networks for Emotion and Gender Classification (arxiv.org)

[9] Ali Ghofrani, Rahil Mahdian Toroghi and Shirin Ghanbari “Realtime Face-Detection and Emotion Recognition Using MTCNN and miniShuffleNet V2” Published in 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI). Realtime Face-Detection and Emotion Recognition Using MTCNN and miniShuffleNet V2 | IEEE Conference Publication | IEEE Xplore

[10] Tee Connie, Mundher Al-Shabi, Wooi Ping Cheah and Michael Goh “Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator” Published in October 2017. Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator | SpringerLink

[11] Minchul Shin, Munsang Kim and Dong-Soo Kwon “Baseline CNN structure analysis for facial expression recognition” Published in: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). Baseline CNN structure analysis for facial expression recognition | IEEE Conference Publication | IEEE Xplore

[12] A. Nasuha, F. Arifin, A. S. Priambodo, N. Setiawan and N. Ahwan “Real Time Emotion Classification Based on Convolution Neural Network and Facial Feature” Citation A Nasuha et al 2021 J. Phys.: Conf. Ser. 1737 012008. Real Time Emotion Classification Based on Convolution Neural Network and Facial Feature - IOPscience

[13] Chengkun Shi, Xiaoqing Zhou, ZhiCheng Zhang, Jie Xu “HM-FER: Hybrid attention mechanism integrate Multiple scales for Facial Expression Recognition” Published on 26 January 2023. HM-FER: Hybrid attention mechanism integrate Multiple scales for Facial Expression Recognition | IEEE Conference Publication | IEEE Xplore

[14] Lutfiah Zahara, Purnawarman Musa, Eri Prasetyo Wibowo, Irwan Karim, Saiful Bahri Musa. ‘‘The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi’’. Pages 2 - 3

[15] P. Ecman, “Micro Expressions,” 2017. [Online].