Deramgozin, M., Jovanovic, S., Rabah, H., & Ramzan, N. (2021). A hybrid explainable AI framework applied to global and local facial expression recognition. In *2021 IEEE International Conference on Imaging Systems and Techniques (IST)* IEEE. https://doi.org/10.1109/IST50367.2021.9651357

"© © 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works."

A Hybrid Explainable AI Framework Applied to Global and Local Facial Expression Recognition

M. Deramgozin, S. Jovanovic, H. Rabah
Institute Jean Lamour
University of Lorraine
Nancy54000, France
mohammad-mahdi.deramgozin@univ-lorraine.fr

N. Ramzan
University of the West of Scotland
Paisley PA1 2BE, United Kingdom
Naeem.Ramzan@uws.ac.uk

***Abstract*—Facial Expression Recognition (FER) systems have many applications such as human behavior understanding, human machine interface, video games and health monitoring. The main advantage of the traditional white box methods is their explainability. However, the accuracy of recognition of these methods is completely reliant on the extracted features. On the other hand, the use of deep neural networks has advantage regarding the overall precision compared to traditional methods. Indeed, they are considered as black box methods and thus suffer from lack of reliability and explainability. In this work, we introduce a hybrid AI explainable framework (HEF) composed of a main functional pipeline comprising a Convolutional Neural Network (CNN) to classify input images and an explainable pipeline using Facial Action Units and application agnostic models LIME providing more useful data allowing to explain the obtained results and reinforce the decision provided by the main functional pipeline. The proposed HEF has been validated on the CK+ dataset and shows very promising results in terms of explainability of the obtained results.**

***Index Terms*—Facial Expression Recognition, Convolutional Neural Networks (CNN), eXplainable Artificial Intelligence (XAI), Emotion classification, Multi Layer Perceptron (MLP)**

I. Introduction

Nowadays, Facial Expression Recognition (FER) is a core part of many systems finding their place in different application fields such as human behavior understanding, human machine interface, video games and health monitoring. This computer vision topic is very challenging and remains a very active research domain. Various machine learning techniques are used to address the FER challenges, with a particular interest in the use of deep learning in recent years.

The early research on FER have relied on the feature extraction followed by classification. The feature extraction aims at getting the distinguishable features for each expression. Methods based on Gabor wavelets were introduced for coding facial expression [1]. Other filters such as Local Binary Patterns (LBP) and pattern descriptor Local Directional Pattern (LDP) have also been used. In the recognition part, classifiers such as Support Vector Machines (SVM), K-Nearest Neighborhood (KNN), or Principal Component Analysis (PCA) classifiers were usually trained with the extracted features from the previous phase [2]–[4]. Various techniques are instigated to enhance features extraction generally by focusing on some areas of the face (e.g. eyes and mouth). The fundamental actions of individual muscles or groups of muscles of the face that play a crucial role in FER systems are called Action Units (AUs). The coordinates of each AU can be obtained by using facial landmarks whose total number is 68. Since Ekman et *al.* [5] came up with the Facial Action Coding System (FACS) to classify emotions based on AUs, this system has become a standard of most FER models for the estimation and recognition of AUs. In the literature, Action units (AU) were mostly used as features, feeding an MLP and KNN classifiers to recognize emotions [6]. The main advantage of these methods is their interpretability. However, the overall accuracy of recognition is totally reliant on the extracted features (read AUs).

Deep learning has been widely exploited the last years in different aspects of FER. For instance, Convolutional Neural Networks (CNN) were used to extract action units in [7]. Similarly, the Convolutional Neural Networks (CNNs) were applied to regions of interest in faces to extract important features [8]. Various CNN models such ResNet and VGG have also been proposed for emotion recognition [9]–[11]. Despite the advantages of CNNs regarding the precision compared to traditional methods, they are considered as black box methods suffering from lack of reliability and interpretability [12]. Due to this fact, eXplainable Artificial Intelligence (XAI) methods such as Grad-CAM [13], SHapley Additive exPlanations (SHAP) [14], Layer-Wise Relevance Propagation (LRP) and Local Interpretable Model-agnostic Explanations (LIME) [15] were introduced. The main goal of these methods is to find and highlight the major parts of the input image that have an effect on the classifier's decision. Even though these methods allow the visualisation of the contributing parts of the image for the classification, the results remain difficult to interpret with regard to facial action units, which are the golden standard in facial emotion recognition. In this paper, we propose a hybrid explainable framework (HEF) composed of two pipelines: the first functional pipeline is the black box approach comprising a CNN to classify input images whereas the second one is an explainable pipeline using Facial Action Units and agnostic LIME model providing more useful data allowing for the visualisation of active regions in the CNN model and thus helping to explain the obtained results. In addition, the proposed framework allows to reinforce (or un-

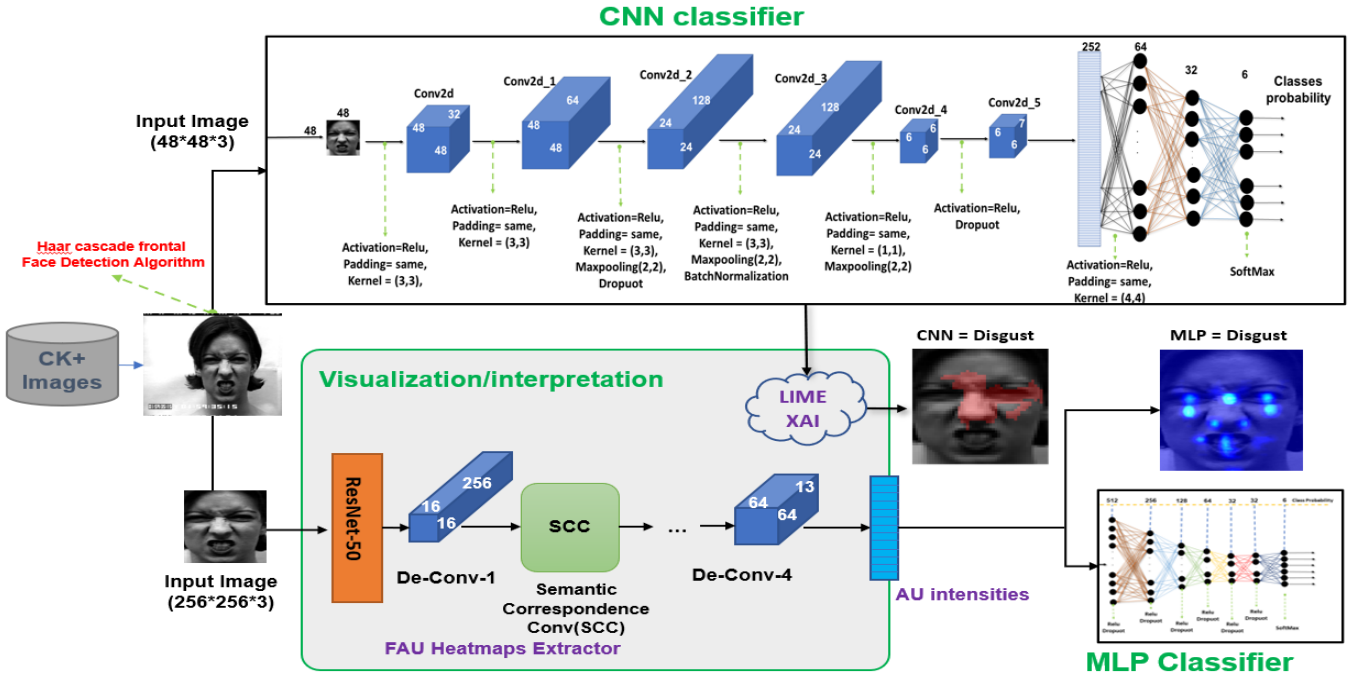

Fig. 1. Hybrid Explainable Framework (HEF): The main diagram of the proposed framework: the first pipeline (CNN, top row) works on the general characteristics of the image whereas the second one (bottom row) is for explainability part and decision reinforcement by using the action units extraction and LIME method. The first pipeline outputs the predicted classes whereas the second one the super pixels and heat maps of the action units on the input image and the predicted classes from the second pipeline.

dermine) the decision provided by the main functional pipeline by reusing the explainable part results and some additional blocks to complete the second functional pipeline. The rest of this paper is organized as follows: Section 2 details the proposed hybrid explainable framework. The obtained results and carried experiments are presented in Section 3. Finally, Section 4 draws some conclusions and perspectives for the future work.

II. Proposed framework

The proposed framework is composed of two main pipelines. The first pipeline is a trained CNN which analyses faces from input images and predict facial recognition classes, whereas the main goal of the second pipeline is to extract features and metrics necessary to interpret the predicted result with the first one. Indeed, in this second pipeline we find a facial action unit extractor and LIME visualisation method. The CK+ [16] is used to train and test both pipelines of the proposed framework. Haarcascade frontal face detection proposed in [17] is used to detect and adapt the face from the original image to remove unnecessary parts for each pipeline. Image resizing is applied to change the size of the input images to 48×48 for the CNN classifier and 256×256 for the facial action unit extractor.

A. Functional pipeline

The main functional pipeline (top row of Fig. 1) is a CNN classifier. This CNN classifier is built from scratch and composed of 6 convolutional layers. The convolutional layers have respectively 32 48×48 filters, 64 3×3 filters, 128 3×3, 128 3×3 filters, 64 3×3 filters, and 34 3×3. A stride of size 1, batch normalization, dropout, max-pooling and ReLU as the activation function are applied. The hidden layer in the first fully connected (FC) layer had 256 neurons, the second FC layer has 64 neurons, whereas the third FC had 32 neurons. Batch normalization, dropout and Softmax are used in FC. The output of fully connected layers is sized to predict 6 emotion classes.

B. Explainable pipeline

The explainable pipeline is composed of three parts: a LIME visualization part which uses the results of the main functional pipeline to identify the superpixels at the origin of the predicted classes; a facial action units extraction parts to identify the Action Units identified for an input image and help the interpretation of both LIME and main function pipeline results; and an MLP classifier which is an additional block reusing the results of the FAU extractor to reinforce (or undermine) the results obtained by the main functional pipeline. In other words, this part can be considered as a redundant functional pipeline built from the available explainable layer.

1) LIME Visualization: To check the reliability of the CNN model, LIME (Local Interpretable Model-Agnostic Explanations) is used to find the superpixels involved in the classifier's decision. LIME trains local surrogates to explain a single prediction using user goal points and their neighborhood. Then, using the sampled points and black-box predictions in this

neighborhood, it trains a weighted intrinsically interpretable surrogate model. Lastly, it interprets the surrogate model [12].

2) Facial Action Unit extractor: The extraction of the intensity of facial action units is based on the model propose in [18]. The FAU part is based on an AutoEncoder using a pre-trained Resnet-50 model as the encoder to extract the action units from the input image, then multiple Semantic Correspondence Convolutional (SCC) and De-convolutional layers performing feature Upsampling and using graph classification via K-Nearest Neighborhood algorithm. This model is modified to capture the intensity of 13 action units in order to support all facial expressions. The 13 AUs needed in this work are: AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU12, AU15, AU17, AU20, AU23 and AU26.

3) Redundant functional pipeline: As explained earlier in this paper, the vast majority of the classical and traditional approaches proposed in the literature are based on the extracted facial features among others we find Facial Action Units. Similarly, in this paper the FAU part is essentially used within the explainable pipeline to help interpretation of the obtained results. However, these already prepared FAUs can be reused to reinforce the main functional pipeline in terms of emotion class predictions. For this, a simple Multi-Layer Perceptron (MLP) model has been added at the output of the FAU extractor. It has been trained with the same prediction classes as the main functional pipeline. The main parameters of this MLP layer are: Seven Fully Connected(FC) layers with 512, 256,128,64,32,32 and 6 neurons as the last layer(corresponding to six emotion classes), Relu activation functions for all FC layers and a Dropout layer after each FC layer to remove inactivated features reductions in settings.

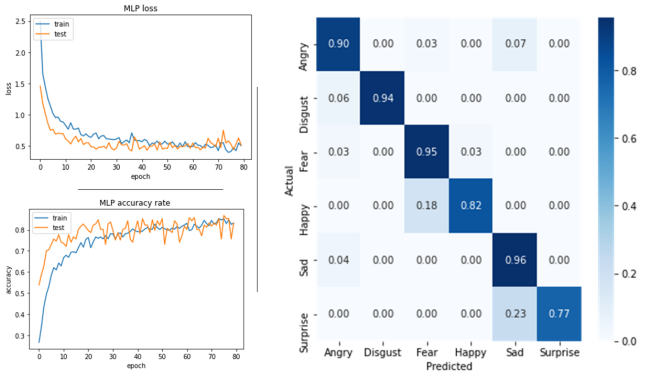

Fig. 2. MLP model (redundant functional pipeline): accuracy rate, loss function and confusion matrix

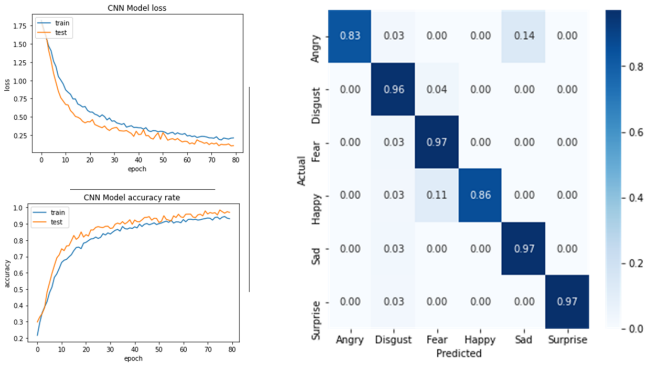

Fig. 3. CNN model (main functional pipeline): accuracy rate, loss function and confusion matrix of CNN model training

III. Results and Discussion

Both CNN and MLP models were trained using 80 epochs via the Adam [19] optimization algorithm, as well as the categorical cross-entropy loss function [20]. During the training and testing phases of both models, overfitting was not observed. The classification reports are shown in Fig. 2 and Fig. 3. In the CNN model, the extracted facial images from the CK+ dataset were used as inputs. To train the FAU extractor model as an auto-encoder, a ground truth containing the coordinates and intensities of the AUs is needed. In this work, the Openface tool [21] was used to extract these information.

To illustrate the proposed framework, an image from the input data set labelled "Disgust" is used for both functional (CNN) and explainable (LIME, FAU+MLP) pipelines respectively. The output of the functional pipeline CNN (class prediction on top of each image) and associated LIME approach in the explainable pipeline (red superpixels superimposed on the input image) are both shown in Fig. 4. Moreover, in the second part of the explainable pipeline comprising the FAU extractor, the output of each action unit was presented as a heat map on top of the input image (the images labelled from AU01 to AU26, see Fig. 5 (a)). The visual comparison between the results provided with the LIME (red superpixels) and the aggregated outputs from all AUs (in blue) for a given input image (here labelled "Disgust") are shown in Fig. 5 (b). The output of the FAU extractor representing the intensity diagram for each AU involved in 6 classes of emotion is shown in Fig. 6.

The results obtained with the proposed HEF approach were compared with state-of-the-art works. These results are shown in Table I, where our work is compared with the results reported in [11], [22], [23], [24] and [25]. The results in Table I clearly show that the proposed hybrid framework is competitive with the state of the art works in terms of overall accuracy. We recall that the main objective of this work is not to obtain the best facial emotion recognition results but to provide a framework allowing to understand, interpret and

Fig. 4. The Superpixels involved in the CNN classifier on the "Disgust" class. The degree of proximity of the input image to each class is given in the image title.

Fig. 5. The predicted classes of the CNN-LIME and FAU-MLP models: a) Original image and 13 AUs heatmaps, as well as the superpixels on the CNN output, b) Visual comparison between the original input, Max AUs heatmaps and CNN-LIME output.

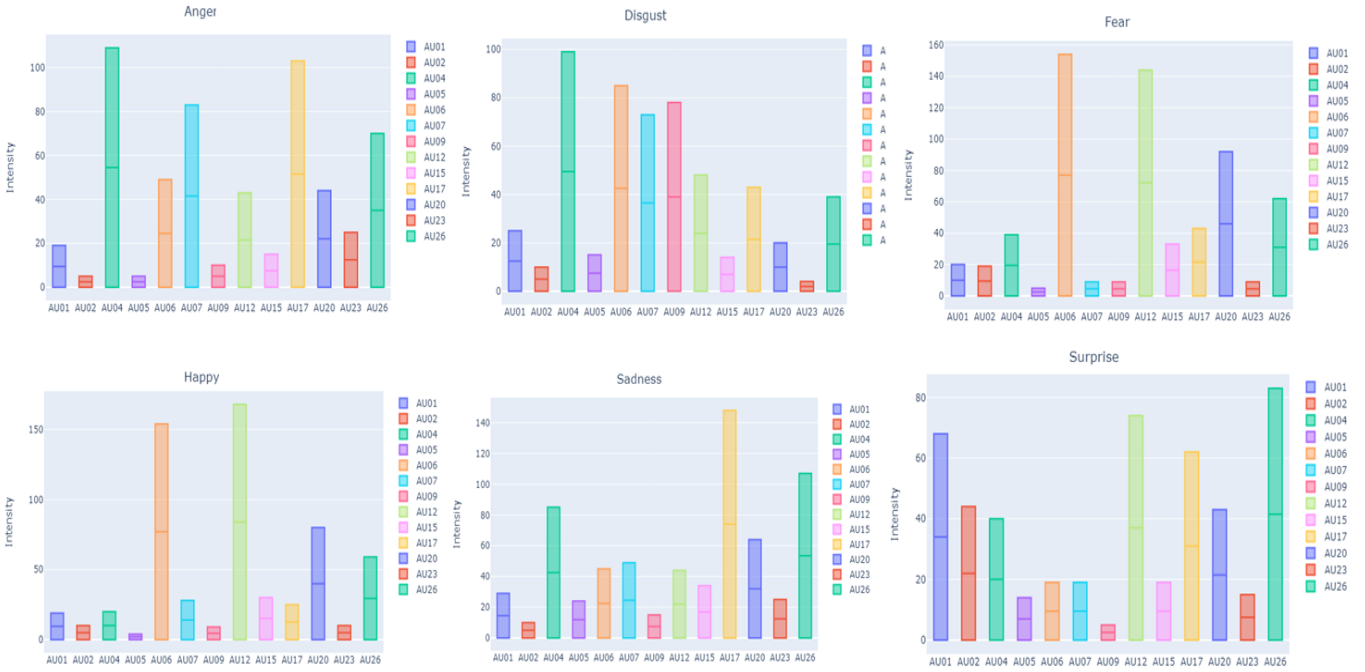

Fig. 6. AUs intensities diagrams for Anger, Disgust, Fear, Happy, Sadness and Surprise classes. The intensity of each AU represents its involvement in the corresponding class.

TABLE I
COMPARISON WITH THE STATES OF THE ART ON CK+ DATASET

Method	ACC
Sun et al [8]	87.20
DIA [11]	89.51
Zahng et al [10]	92.73
Elfatih et al [26]	92.73
Shahid et al [27]	94.90
SACNN-LSTM [28]	95.15
eXnet [29]	96.75
Ours (FAU+MLP)	88
Ours (CNN)	**97.03**

gain confidence in the obtained results with the functional pipeline. Consequently, any CNN-based approach providing better results in terms of accuracy can be used in the proposed framework instead of the employed 6-layer CNN architecture. On the other hand, the results in terms of accuracy obtained with the redundant functional pipeline reusing the explainable outputs (the FAU extractor outputs) as inputs of an additional MLP layer are also shown in Table I. As expected, they are lower than the ones obtained with the main functional pipeline, but can help reinforce its final decision.

In Fig. 6, as mentioned earlier, the intensity diagram for each AU involved in 6 classes of emotion is shown. This intensity diagram of AUs extracted in this work is consistent with the 6 basic emotions reported in the seminal work presented in [5]. However, there are a small number of differences in some cases from the used CK+ dataset which are mostly due to the inter-subject variability and in the difficulty to express clearly only one emotion in terms of FACS [5] during the dataset building without introducing some unwanted artefacts (i.e. exaggerated mouth stretch, etc). This was reported in [30] where also a list of compound emotional categories (i.e. happily surprised, fearfully angry, etc) have been presented allowing to explain the presence of some at first sight unexpected AUs in the obtained results.

IV. CONCLUSION

In this work, a hybrid AI explainable framework composed of a functional and an explainable pipeline were used to both recognize and classify the Facial Expressions of input facial

images and to provide additional information for the interpretation and understanding of the obtained results. The functional pipeline comprises a 6-layer CNN architecture and allows to output the 6 basic emotion categories. It was backed with an explainable layer comprising a FAU extraction module whose outputs are crucial in interpretation and understanding of the obtained results with the main functional flow. Interpretability is a very important part of this hybrid approach where by extracting the action units additional useful information can help gaining in confidence of the obtained results provided with the main functional pipeline. Moreover, as presented in this approach, this explainable support can also be used as a reinforcement of the main decision pipeline. As perspectives of this work, we plan to enlarge the list of extracted AUs and to take into consideration the inter-subject variability and compound emotion categories. Thus, the better interpretation of the "difficult" facial images can be obtained. Moreover, other methods like Graph Convolutional Networks (GCN) [31] may be considered to improve the FAU extraction part of this framework as well as the possibilities to transpose the proposed framework to other application domains than facial expression recognition.

REFERENCES

[1] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, *Coding Facial Expressions with Gabor Wavelets*, vol. 1998. May 1998. ECC: 0002361 Journal Abbreviation: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, April Pages: 205 Publication Title: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, April.

[2] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip movement synthesis from speech based on Hidden Markov Models1Speech files available. See http://www.elsevier.nl/locate/specom.1," *Speech Communication*, vol. 26, pp. 105–115, Oct. 1998. ECC: No Data (logprob: -324.763) 61 citations (Crossref) [2021-06-22].

[3] L. Yao, Y. Wan, H. Ni, and B. Xu, "Action unit classification for facial expression recognition using active learning and SVM," *Multimedia Tools and Applications*, Apr. 2021. ECC: 0000000 0 citations (Crossref) [2021-05-17].

[4] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu, "A principal component analysis of facial expressions," *Vision Research*, vol. 41, pp. 1179–1208, Apr. 2001. 276 citations (Crossref) [2021-06-22] ECC: 0000552.

[5] P. Ekman and W. V. Friesen, *Facial action coding system: manual*. Palo Alto, Calif.: Consulting Psychologists Press, 1978. OCLC: 5851545.

[6] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science*, vol. 108, pp. 1175–1184, Jan. 2017. 66 citations (Crossref) [2021-07-01] ECC: 0000127.

[7] C. Pramerdorfer and M. Kampel, "Facial Expression Recognition using Convolutional Neural Networks: State of the Art," *arXiv:1612.02903 [cs]*, Dec. 2016. ECC: 0000124 arXiv: 1612.02903.

[8] X. Sun, S. Zheng, and H. Fu, "ROI-Attention Vectorized CNN Model for Static Facial Expression Recognition," *IEEE Access*, vol. 8, pp. 7183–7194, 2020. 7 citations (Crossref) [2021-06-29] ECC: 0000006 Conference Name: IEEE Access.

[9] K. Weitz, T. Hassan, U. Schmid, and J.-U. Garbas, "Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods," *tm - Technisches Messen*, vol. 86, pp. 404–412, July 2019. Publisher: De Gruyter Oldenbourg Section: tm - Technisches Messen 9 citations (Crossref) [2021-04-16].

[10] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–Temporal Recurrent Neural Network for Emotion Recognition," *IEEE Transactions on Cybernetics*, vol. 49, pp. 839–847, Mar. 2019. 57 citations (Crossref) [2021-06-29] ECC: 0000125 Conference Name: IEEE Transactions on Cybernetics.

[11] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive Into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition," p. 10.

[12] S. Masís, *Interpretable Machine Learning with Python*. Packt Publishing. ECC: 0000001.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, pp. 336–359, Feb. 2020. ECC: 0000364 200 citations (Crossref) [2021-05-01].

[14] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *arXiv:1705.07874 [cs, stat]*, Nov. 2017. arXiv: 1705.07874.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," *arXiv:1602.04938 [cs, stat]*, Feb. 2016. ECC: 0004671 arXiv: 1602.04938 version: 1.

[16] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, (San Francisco, CA, USA), pp. 94–101, IEEE, June 2010. 1297 citations (Crossref) [2021-04-16].

[17] S. Mehtab and J. Sen, *Face Detection Using OpenCV and Haar Cascades Classifiers*. Mar. 2020. ECC: No Data (logprob: -57.959).

[18] Y. Fan, J. Lam, and V. Li, "Facial Action Unit Intensity Estimation via Semantic Correspondence Learning with Dynamic Graph Convolution," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12701–12708, Apr. 2020. 0 citations (Crossref) [2021-04-30] Number: 07.

[19] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017. ECC: 0002750 arXiv: 1412.6980.

[20] Z. Zhang and M. Sabuncu, "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels," p. 11. ECC: 0000393.

[21] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, (Xi'an), pp. 59–66, IEEE, May 2018. 204 citations (Crossref) [2021-04-30].

[22] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution," *arXiv:1608.01041 [cs]*, Sept. 2016. ECC: 0000216 arXiv: 1608.01041.

[23] C. Huang, "Combining convolutional neural networks for emotion recognition," in *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pp. 1–4, Nov. 2017. ECC: 0000015 7 citations (Crossref) [2021-06-29].

[24] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion Recognition in Speech using Cross-Modal Transfer in the Wild," *arXiv:1808.05561 [cs]*, Aug. 2018. ECC: 0000109 arXiv: 1808.05561.

[25] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition," *arXiv:1905.04075 [cs]*, Sept. 2019. ECC: 0000005 arXiv: 1905.04075.

[26] E. E. Mustafa and G. Z. A. Salh, "FACIAL EMOTION RECOGNITION BASED ON DEEP LEARNING TECHNIQUE," . *Vol.*, no. 4, p. 12, 2021. ECC: 0000002.

[27] A. R. Shahid, S. Khan, and H. Yan, "Contour and region harmonic features for sub-local facial expression recognition," *J. Vis. Commun. Image Represent.*, 2020. 0 citations (Crossref) [2021-06-29] ECC: 0000001.

[28] J. Liu, Y. Feng, and H. Wang, "Facial Expression Recognition Using Pose-Guided Face Alignment and Discriminative Features Based on Deep Learning," *IEEE Access*, vol. 9, pp. 69267–69277, 2021. ECC: No Data (logprob: -251.89) Conference Name: IEEE Access.

[29] M. N. Riaz, Y. Shen, M. Sohail, and M. Guo, "eXnet: An Efficient Approach for Emotion Recognition in the Wild," *Sensors*, vol. 20, p. 1087, Jan. 2020. ECC: 0000005 6 citations (Crossref) [2021-06-29] Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[30] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[31] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph Convolutional Networks for Hyperspectral Image Classification," *arXiv:2008.02457 [cs]*, Jan. 2021. ECC: 0000000 arXiv: 2008.02457.