Facial Emotion Recognition Using Ensemble Learning

GuanQun Xu and Wei Qi Yan
Auckland University of Technology, 1010 New Zealand

ABSTRACT

Facial Emotion Recognition (FER) is the task of identifying human emotions from facial expressions. Most CNN networks are capable of recognizing expressions relatively accurately, especially large CNN networks. It is conducive to the recognition of expressions when more features are extracted from the target. The purpose of this book chapter is to improve accuracy of facial emotion recognition using integrated learning of lightweight networks without increasing the complexity or depth of the network. Compared to single lightweight models, it made a significant improvement. For a solution, we proposed an ensemble of mini-Xception models, where each expert is trained for a specific emotion and lets confidence score for the vote. Therefore, the expert model will transform the original multiclass task into binary tasks. We target the model to differentiate between a specific emotion and all others, facilitating the learning process. The principal innovation lies in our confidence-based voting mechanism, in which the experts "vote" based on their confidence scores rather than binary decisions. Furthermore, while we found the imbalance between emotion datasets, we introduced data augmentation methods, through oversampling positive samples to improve training effectiveness. Contrasted with the conventional mini-Xception model, our ensemble learning method showcased superior robustness, especially in ambiguous scenarios.

Keywords: Facial Emotion Recognition, mini-Xception, Ensemble learning, TensorFlow

INTRODUCTION

Facial Emotion Recognition (FER) currently lying at the crossroads of psychology and computer science, has grown immensely with the advent of machine learning and more specifically in deep learning. Historically, understanding and interpreting human emotions were subjective, relying heavily on human intuition and judgment. However, with the increasing integration of technology into our daily lives, objective identification of emotions through machines becomes not only desirable but, in many scenarios, it is essential.

Delving into the nature of emotions, the basic emotion posits that humans universally experience the foundational emotions, namely, happiness, sadness, fear, anger, disgust, and surprise. These fundamental emotional states can be seen as building blocks, from which more nuanced emotions—such as fatigue, anxiety, or satisfaction—emerge. Practical applications of FER are vast and varied. In human-computer interaction, deep learning algorithms can be designed to adapt and respond based on the user's emotional state, creating a more intuitive and empathetic user experience. In healthcare, it can be employed for monitoring patients for signs of pain or distress, especially if they cannot communicate verbally. In automotive industry, FER can be used to

monitor driver's emotions and alertness, potentially preventing accidents caused by drowsiness or distress.

In 2006, Hinton introduced the ground-breaking theory of deep learning and subsequently applied it innovatively to image processing. Deep learning, fundamentally rooted in the deep neural network, is a specialized subset of artificial neural networks. The foundation of deep learning is established upon the research progress in artificial neural networks. By adjusting the number of hidden layers, one can derive an artificial neural network model with multiple hidden layers. Hidden neural networks are able to learn more effectively, mirroring the cognitive processes of the human brain. This facilitates the efficient extraction of image features (Feng et al., 2020).

Among deep learning architectures, CNNs became the poster child for FER. They consist of convolutional layers that can automatically and adaptively learn spatial hierarchies of features from input images. This property alleviated the need for hand-crafted features, a limitation of traditional methods. Layers within CNNs, such as pooling layers, helped in reducing spatial dimensions while retaining crucial information. Activation functions introduced nonlinearity, while enabling the network to capture complex relationships.

The mini-Xception model draws inspiration from the original "Xception" architecture, which stands for "Extreme Inception" (Li et al., 2022). In the Keras deep learning library, Xception was designed to improve upon the Inception architecture by using depthwise separable convolutions. The result of mini-Xception model comprises of four depthwise separable convolution blocks. The batch normalization processes the output to stabilize and accelerate the training process. This is complemented by the introduction of the ReLU activation function, which infuses the model with the necessary non-linearity. In the forward pass, the SoftMax function is invoked to facilitate multi-class classification of the results.

In the vast landscape of ensemble learning, the idea of leveraging multiple models to make a collective decision is central. One of the most intuitive and widely employed methods to achieve this consensus is through voting mechanisms. The intrinsic capability of deep neural networks to capture intricate patterns means that even a simple procedure like unweighted averaging can significantly enhance performance. By averaging across multiple networks, one can effectively reduce the model variance. This is especially impactful given that deep artificial neural networks (ANNs) are characterized with high variance but low bias. If the underlying models are sufficiently diverse or uncorrelated, their collective variance can be markedly diminished if averaged.

In hard voting, each model in the ensemble "votes" for a specific class. The class that receives the majority of votes is chosen as the final prediction. It's straightforward and doesn't require probability estimates. The advantages of voting mechanisms is by aggregating predictions, the ensemble smoothens out the biases and variances of individual models, which leads to a model that's less prone to overfitting.

In the context of ensemble models, majority voting refers to taking the mode of predictions across all models to arrive at the final prediction. This method is often effective if all models are equally reliable.

$$y_{\text{final}} = mode(y_1, y_2, \dots, y_7) \tag{1}$$

where y_i is the prediction from the i^{th} expert model. If most models predict a specific expression for an input image, that expression is taken as the final predicted class.

Instead of giving equal importance to all models, weighted voting takes into consideration of the confidence or reliability of each prediction of model. This ensures that more reliable models have a greater influence on the final decision.

$$y_{\text{final}} = argmax_j \sum_{i=1}^{7} w_i p(y = j \mid x, M_i) \quad (2)$$

where $p(y = j \mid x, M_i)$ is the probability of data x belonging to class *j* as predicted by model M_i, w_i is the weight (or confidence score) of the i^{th} model. $argmax_j$ ensures that the class with the highest aggregated score is selected as the final prediction.

Confidence scores play a crucial role in our voting mechanism. Pertaining to a binary classification, where each expert model determines whether an input image belongs to a specific expression or the "other" class:

$$c_i = p(y = 1 \mid x, M_i) \quad (3)$$

where c_i is the confidence score for the i^{th} model, $p(y = 1 \mid x, M_i)$ is the probability that data x belongs to the target expression as predicted by model M_i. High confidence scores indicate strong certainty in a prediction of model, making it a suitable weight in our weighted voting scheme.

In this book chapter, a novel approach is proposed for emotion recognition by leveraging a unique methodology that incorporates weighted confidence scores. The primary innovation lies in the dynamic incorporation of predefined weights to the confidence scores produced by individual models specialized in recognizing specific emotions. These weights are not arbitrary but are derived from prior accuracy metrics, offering a degree of reliability and prediction. Based on historical accuracy rates, weights are assigned to each emotion. For instance, "Anger" is associated with a weight of 0.79. This signifies that the prediction confidence of this model for "Anger" would be adjusted by multiplying it with 0.79.

For each label or emotion, the corresponding specialized model is utilized to predict the confidence score for that particular emotion on the test image data. Each raw confidence score is then multiplied by its respective weight, creating a weighted confidence score. This step is central to this approach. The emotion associated with the highest weighted confidence score is considered the predicted emotion for the test image.

RELATED WORK

A facial expression recognition method was proposed based on convolutional neural network ensemble learning (Jia et al. 2020). The method takes use of three sub-networks and an SVM classifier to integrate the output of the three networks to obtain the final result. The model achieved a facial expression recognition accuracy of 71.27% on the FER2013 dataset.

The mini-Xception model – a lightweight and efficient variant of the original Xception architecture was designed for on-device real-time applications. Its depth-wise separable convolutions ensure fewer parameters and operations without compromising much on performance, which makes it a prime candidate for ensemble learning, especially in scenarios demanding speed and efficiency.

In Facial Emotion Recognition (FER), fine distinctions between emotions often mean the difference between accurate and subpar models, the amalgamation of ensemble learning and mini-Xception model becomes particularly compelling. By pooling together multiple mini-Xception models, each was trained with slight variations or focuses, the ensemble learning is positioned to capture a broader range of facial emotional nuances.

To balance the classes, we artificially increase the number of instances of the minority class by using methods like SMOTE, which generates synthetic samples; Brightness & contrast adjustment is a method for modifying the brightness and contrast of images by simulating different lighting conditions, crucial for FER in diverse environments; Elastic deformations are designed for slight warping of facial images that can mimic different facial expressions and nuances.

$$N' = N + SMOTE\left(N_{\text{minority}}, \alpha\right) \tag{4}$$

where N' is the new sample size after oversampling. N is the original sample size. N_{minority} represents the count of minority class samples. α is the oversampling ratio, determining how many synthetic samples to create. Instead of increasing the minority class, undersampling reduces the instances of the majority class to balance the classes.

$$N' = N - \text{RandomUndersample}\left(N_{\text{majority}}, \beta\right) \tag{5}$$

where N' is the new sample size after undersampling, N is the original sample size, N_{minority} represents the count of minority class samples, β is the undersampling ratio, dictating the fraction of majority class samples to be removed.

Binary mini-Xception models are at the heart of this ensemble learning, each fine-tuned for recognizing a specific emotion. Instead of a singular model attempting to classify multiple emotions, this approach offers dedicated experts for each emotion, ensuring a deeper understanding and more precise classification for that particular emotion.

Based on confidence-based voting mechanism, once each binary mini-Xception model makes a prediction for its corresponding emotion, the system doesn't merely count the votes. Instead, it considers the confidence or probability associated with each prediction. This approach ensures that the final decision takes into account not just the number of models favouring a particular emotion but also their respective certainty levels.

THE METHODOLOGY

In this book chapter, we examine the lightweight model mini-Xception. We used Jupyter Lab and Colab for code development, experimentation, and visualization. Jupyter Lab was used on an operating system based on Ubuntu 20.04 LTS, which was equipped with an Nvidia 3070 GPU.

With TensorFlow running on Google Colab, this project was able to use its cloud-based computational resources in parallel and to utilize its integrated T4 GPUs. A flexible computational environment was achieved by combining local resources with cloud-based platforms. The model was trained using Jupyter Lab locally following stability and model debugging using Colab.

Our dataset of FER2013 is based on an automatically collected dataset derived from the Google Image API. In the FER-2013 dataset, human accuracy averages around 65%. However, Tang's approach in 2013 led to a test accuracy of 71.2% by utilizing a CNN combined with L2-SVM loss, marking a significant milestone and subsequently winning him the ICML 2013 Challenges in Representation Learning (Tang, 2013). Subsequent advancements have yielded even better results. For instance, in 2016, Kim et al. reported an impressive 73.73% test accuracy. The methodology involved an ensemble of CNNs processing both aligned and non-aligned images. A standout aspect of the method was a pre-processing alignment step conducted by using a dedicated Deep Convolutional Network, which essentially learned from an optimal mapping (Kim, Dong, Roh, Kim, & Lee, 2016). Similarly, in 2017, Connie proposed a unique model blending both SIFT and CNN features. The innovative approach, which aggregated insights from three distinct models, resulted in a commendable 73.4% accuracy (Connie, Al-Shabi, Cheah, & Goh, 2017). Notably, Zhang et al. achieved a 75.1% test accuracy by amalgamating training data from a variety of sources, underlining the potential of harnessing diverse information in training (Zhang, Luo, Loy, & Tang, 2015).

The FER2013 dataset suffers from an imbalance in the distribution of its emotion classes. For the applications, such an imbalance can lead to a biased model that disproportionately favourites the majority class. In response to this, we implemented a hybrid technique to level the class distribution. (Renda et al., 2019)

We expanded the representation of the target emotion class by replicating its instances. Specifically, the samples of the desired emotion (denoted as the target emotion) were oversampled by using a factor (*pos_multiplier)*, effectively increasing their count. For instance, with *pos_multiplier* is set to 2, the target class samples would be doubled. To further refine the balance, the non-target emotion classes were undersampled. This was achieved by randomly removing a proportion (*neg_multiplier*) of samples from these classes. For example, a *neg_multiplier* such as 0.3 would result in the removal of approximately 30% of the non-target class samples.

Recognizing the notably smaller representation of the “Disgust” emotion in the dataset, we adjusted the oversampling factor specifically for this class, thereby with a greater emphasis during the balancing process. After postprocessing, the balanced dataset underwent a re-evaluation of its class distribution. The resultant dataset exhibited a more equitable distribution between the target emotion and the remaining emotions, with a sample ratio of approximately 1:1.5 (target to non-target). The balanced datasets can facilitate more robust model training, fostering improved generalization and reduced bias towards any specific emotion class.

In the context of our facial emotion recognition, each emotion-specific model, termed as an “expert”, focuses on the accurate detection of a particular emotion. Each expert was individually trained and validated against the FER2013 dataset. Each expert model underwent an initial training phase of 50 epochs. Post this phase, the models were equipped with pre-trained weights, and the

training process was extended for an additional 20 epochs. For most emotion categories, the models showcased impressive accuracy, often hovering around the 85% mark in binary classification scenarios.

RESULT ANALYSIS

(a)

	precision	recall	f1-score	support
Anger	0.74	0.68	0.71	495
Disgust	0.20	0.36	0.26	50
Fear	0.76	0.63	0.69	514
Happy	0.93	0.93	0.93	891
Sad	0.71	0.72	0.72	608
Surprise	0.87	0.84	0.86	406
Neutral	0.72	0.83	0.77	625
accuracy			0.78	3589
macro avg	0.71	0.71	0.70	3589
weighted avg	0.79	0.78	0.78	3589

(b)

	precision	recall	f1-score	support
Anger	0.66	0.64	0.65	101
Disgust	0.36	0.33	0.35	12
Fear	0.62	0.70	0.66	101
Happy	0.93	0.93	0.93	189
Sad	0.70	0.53	0.61	131
Surprise	0.84	0.86	0.85	87
Neutral	0.68	0.81	0.74	97
accuracy			0.75	718
macro avg	0.69	0.69	0.68	718
weighted avg	0.75	0.75	0.75	718

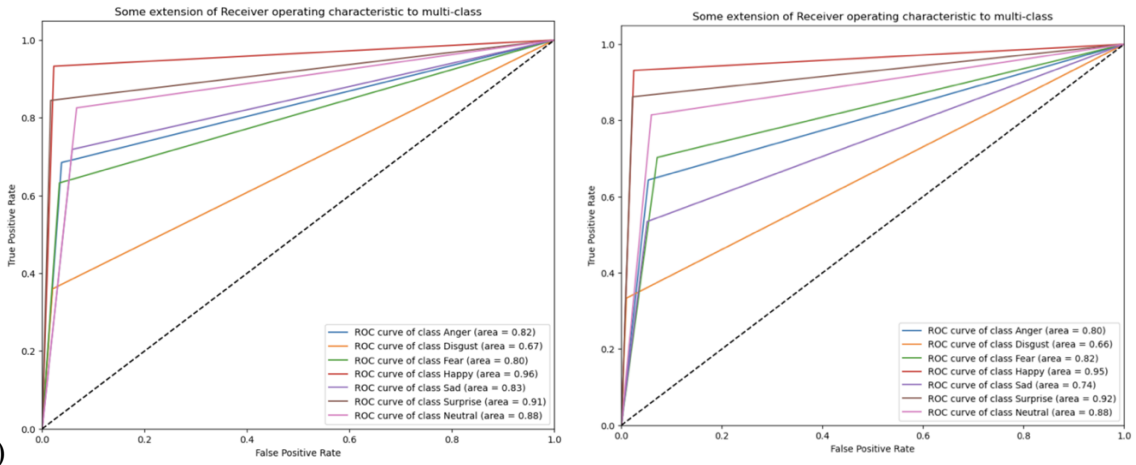
Figure 1: Comparison between (a) If all confidence scores = 1.0 and (b) If confidence scores change

From Figure 1 (a), with an overarching accuracy of 78%, the model demonstrates considerable strength in recognizing a broad spectrum of emotions. Notably, its prowess is pronounced in distinguishing emotions like “Happy” and “Surprise”, as evidenced by the remarkable F1 scores of 0.93 and 0.86, respectively. Moreover, the model adaptiveness extends to the “Neutral” emotion, reflected by a noteworthy recall rate of 0.83. This metric signifies the ability of model to effectively capture the majority of instances associated with neutral expressions, underscoring its balanced performance across various emotion categories.

Figure 1 (b) shows the performance metrics after changing the confidence score for each emotion category: “Anger” at 1.0, “Disgust” at 0.8, “Fear” at 1.0, “Happy” at 1, “Sad” at 0.91, “Surprise” at 1.0, and “Neutral” at 0.92.

As a result of modifying these confidence scores, there is a discernible shift in performance metrics. By retaining its standard, the model exhibits an overall accuracy of 75%, affirming its consistency in identifying emotions spanning various categories. Notably, its proficiency in detecting “Happy” emotions remains undiminished, as evidenced by an F1 score of 0.93.

As a result of Figure 1 (a) to Figure 1 (b), an enhancement in recall is evident for categories like “Fear” and “Neutral”, registering values of 0.70 and 0.81 respectively, suggesting an improved capacity in accurately detecting true instances of these emotions. However, precision in detecting “Anger” sees a decrease, registering at 0.66, hinting at potential misclassifications where non-anger instances might be erroneously recognized as anger. The emotion “Disgust” emerges as a persistent challenge, yielding a modest F1 score 0.35. The outcome, coupled with diminished support, insinuates a possible paucity of data samples for this class, which may be a contributory factor to its suboptimal performance. Additionally, the “Sad” class reveals a decline in its F1 score to 0.61, attributed to a decrease in recall, suggesting potential oversights in identifying genuine instances of sadness following confidence adjustment. While these confidence alterations render enhancements in specific facets, they also spotlight potential zones in the model which might benefit from meticulous fine-tuning or supplementary data for performance amelioration.



(a) (b)

Figure 2: ROC curve Comparison between (a) If all confidence scores = 1.0 and (b) If confidence scores change

In Figure 2 (a), "Happy" emotion stands out with an impressive AUC (Area Under the Curve) of 0.96, underscoring the exemplary capability to discern true positives from negatives within this category. Meanwhile, emotions such as "Fear", "Sad", "Surprise", and "Neutral" exhibit AUC values surpassing 0.80, reflecting their substantial discriminatory efficacy within the model. However, the "Disgust" class presents a challenge, registering an AUC of 0.67.

Followed the adjustment in confidence scores as shown in Figure 2 (b), there were discernible, albeit minor, shifts in the AUC (Area Under the Curve) values of most emotions. For instance, the AUC for "Fear" witnessed an enhancement, climbing from 0.80 to 0.82. Contrarily, "Sad" experienced a decline, receding from 0.83 to 0.74.

Remarkably, the emotions displayed a robust performance: "Happy" managed to sustain its commendable AUC, registering only a slight dip to 0.95. Meanwhile, "Surprise" showed a promising upward trend, elevating from 0.91 to 0.92. However, challenges persisted with the "Disgust" class.

CONCLUSION

In this book chapter, in advancing emotion recognition through the integration of ensemble learning into the mini-Xception framework, we observed marked enhancements in terms of both accuracy and prediction confidence. Yet, like all empirical investigations, our research is not without its limitations. Here we outline the principal constraints of our study:

Our primary data source was the FER2013 dataset. While it offers a comprehensive collection of facial expressions, potential biases, noise, or imbalances in specific emotional categories may exist. Such shortcomings can influence the ability of model to generalize effectively across a myriad of real-world situations.

While the mini-Xception model offers computational efficiency, its relatively simpler architecture might not encapsulate all subtleties of facial expressions as proficiently as more intricate models.

Furthermore, our ensemble approach, which amalgamates a number of "expert" models, might compromise on real-time processing capabilities due to increased computational demands.

Our ensemble framework fundamentally rests on binary classifiers, dedicated to distinguishing between two emotions. This binary emphasis might not encapsulate the multifaceted nature of human emotions, especially in scenarios where emotions blur boundaries.

REFERENCES

Ahonen, T., Hadid, A., & Pietikäinen, M. (2006). Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(12), 2037–2041. https://doi.org/10.1109/tpami.2006.244

Connie, T., Al-Shabi, M., Cheah, W. P., & Goh, M. K. O. (2017). Facial expression recognition using a hybrid CNN–SIFT aggregator. In Lecture Notes in Computer Science (pp. 139–149). https://doi.org/10.1007/978-3-319-69456-6_12

Chirra, V. R. R., Reddy, U. S., & Kolli, V. K. K. (2021). Virtual facial expression recognition using deep CNN with ensemble learning. Journal of Ambient Intelligence and Humanized Computing, 12(12), 10581–10599.

Cui, W., Yan, W. (2016) A scheme for face recognition in complex environments. *International Journal of Digital Crime and Forensics* (IJDCF) 8 (1), 26-36.

Fan, Y., Lam, J. C., & Li, V. O. K. (2018). Multi-region ensemble convolutional neural network for facial expression recognition. In Lecture Notes in Computer Science (pp. 84–94). https://doi.org/10.1007/978-3-030-01418-6_9

Fayek, H. M., Lech, M., & Cavedon, L. (2016). Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles *vs* soft labels. International Joint Conference on Neural Networks (IJCNN).https://doi.org/10.1109/ijcnn.2016.7727250

Feng, Y., Pang, T., Li, M., & Guan, Y. (2020). Small sample face recognition based on ensemble deep learning. Chinese Control and Decision Conference (CCDC).

Gao, X., Nguyen, M., Yan, W. (2021) Face image inpainting based on generative adversarial network. International Conference on Image and Vision Computing New Zealand.

Gao, X., Nguyen, M., Yan, W. (2022) A face image inpainting method based on autoencoder and adversarial generative networks. Pacific-Rim Symposium on Image and Video Technology.

García, S., & Herrera, F. (2009). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. Evolutionary Computation, 17(3), 275–306. https://doi.org/10.1162/evco.2009.17.3.275

Habib, A. S. B., & Tasnim, T. (2020). An ensemble hard voting model for cardiovascular disease prediction. International Conference on Sustainable Technologies for Industry 4.0 (STI). https://doi.org/10.1109/sti50764.2020.9350514

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. http://export.arxiv.org/pdf/1502.03167

Jia, C., Li, C. L., & Zhou, Y. (2020). Facial expression recognition based on the ensemble learning of CNNs. IEEE International Conference on Signal Processing. https://doi.org/10.1109/icspcc50002.2020.9259543

Kim, B., Dong, S., Roh, J., Kim, G., & Lee, S. Y. (2016). Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). https://doi.org/10.1109/cvprw.2016.187

Kyerematen-Boateng, H., Josyula, D. P., & Conn, M. (2023). Computing confidence score for neural network predictions from latent features. International Conference on Control, Communication and Computing (ICCC). https://doi.org/10.1109/iccc57789.2023.10165294

Le, H., Nguyen, M., Yan, W. Q., & Lo, S. (2021). Training a convolutional neural network for transportation sign detection using synthetic dataset. 2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ). https://doi.org/10.1109/ivcnz54163.2021.9653398

Li, C., Li, D., Zhao, M., & Li, H. (2022). A light-weight convolutional neural network for facial expression recognition using Mini-Xception neural networks. IEEE International Conference on Current Development in Engineering and Technology (CCET). https://doi.org/10.1109/qrs-c57518.2022.00104

Liu, K., Zhang, M., & Pan, Z. (2016). Facial expression recognition with CNN ensemble. International Conference on Cyberworlds.

Liu, M., Yan, W. (2022) Masked face recognition in real-time using MobileNetV2. ACM ICCCV.

Nguyen, M., Yan, W. (2022) Temporal color-coded facial-expression recognition using convolutional neural network. International Summit Smart City 360°: Science and Technologies for Smart Cities.

Nguyen, M., Yan, W. (2023) From faces to traffic lights: A multi-scale approach for emotional state representation. IEEE International Conference on Smart City.

Powers, D. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (Cornell University). https://doi.org/10.48550/arxiv.2010.16061

Renda, A., Barsacchi, M., Bechini, A., & Marcelloni, F. (2019). Comparing ensemble strategies for deep learning: An application to facial expression recognition. Expert Systems with Applications, 136, 1–11. https://doi.org/10.1016/j.eswa.2019.06.025

Song, C., He, L., Yan, W., Nand, P. (2019) An improved selective facial extraction model for age estimation. International Conference on Image and Vision Computing New Zealand.

Sridhar, K., Lin, W., & Busso, C. (2021). Generative approach using soft-labels to learn uncertainty in predicting emotional attributes. International Conference on Affective Computing and Intelligent Interaction (ACII).

Sun, L., Ge, C., & Zhong, Y. (2021). Design and implementation of face emotion recognition system based on CNN Mini-Xception Frameworks. Journal of Physics, 2010(1), 012123. https://doi.org/10.1088/1742-6596/2010/1/012123

Tang, J., Su, Q., Su, B., Fong, S., Cao, W., & Gong, X. (2020). Parallel ensemble learning of convolutional neural networks and local binary patterns for face recognition. Computer Methods and Programs in Biomedicine, 197, 105622.

Tang, Y. (2013). Deep learning using linear support vector machines. (Cornell University).

Wang, H., Yan, W. (2022) Face detection and recognition from distance based on deep learning. Aiding Forensic Investigation Through Deep Learning and Machine Learning Framework. IGI Global.

Wang, Y., & Lu, F. (2021). An adaptive boosting algorithm based on weighted feature selection and category classification confidence. Applied Intelligence, 51(10), 6837–6858. https://doi.org/10.1007/s10489-020-02184-3

Webb, G. I., & Zheng, Z. (2004). Multistrategy ensemble learning: Reducing error by combining ensemble learning methods. IEEE Transactions on Knowledge and Data Engineering, 16(8), 980–991. https://doi.org/10.1109/tkde.2004.29

Yan, W. (2021). Computational methods for deep learning. Spring. https://doi.org/10.1007/978-3-030-61081-4

Zehra, N., Azeem, S. H., & Farhan, M. (2021). Human activity recognition through ensemble learning of multiple convolutional neural networks. Annual Conference on Information Sciences and Systems (CISS). https://doi.org/10.1109/ciss50987.2021.9400290

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2015). Learning social relation traits from face images. IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2015.414