Journal of Methodical Computing and Communications **(JMCC)**, Vol. 2025, pp. 33-40, 2025

Real-time Face Emotion Recognition Using Deep Learning with Two-Tier Architectures for Age and Gender

Dheyaa Saifulislam Ali Al-Obaidi[1]*, Sarah Alissa Mohd Dali[2]

[1] Computer Center, University of Mosul, Mosul, Iraq
[2] Faculty of Information Science & Engineering, Management & Science University, Shah Alam, Malaysia
***Corresponding email:** diyaa.saifalislam@uomosul.edu.iq

Received 18 May 2025; **Revised** 30 June 2025; **Accepted** 15 July 2025; **Published** 3 August 2025

Abstract: The automated recognition of human emotion, age, and gender from facial images is a significant area of research with applications in fields like security, healthcare, and human-computer interaction. While numerous systems exist for these tasks, their accuracy often falls short of satisfactory levels, and identifying robust methods remains a challenge. This study proposes a novel deep learning approach using a two-tier architecture that combines a Convolutional Neural Network (CNN) for emotion recognition with a Local-Deep Neural Network (LDNN) for age and gender classification. The model was trained and tested on the AffectNet and UTKFace datasets, demonstrating high efficacy in both training and real-time modes. The system successfully identifies six basic emotions (happy, sad, anger, fear, disgust, surprise), eight age ranges, and two genders. Our results show a significant improvement in emotion recognition accuracy over preceding studies, validating the effectiveness of the proposed architecture.

Keywords: Convolutional Neural Network, Local-Deep Neural Network, Deep Learning, Face Recognition, Emotion Recognition

1. Introduction

Over the past decade, emotion recognition has become a popular research topic within the computer vision community. It has been applied in various fields due to technological advancements [1]. Moreover, emotion recognition has also been applied in many fields, such as medicine, security, and business [2] However, this study focuses on the field of security. There are many complications in developing an effective real-time emotion recognition system for the human face. The human face is unique and contains vital features that can convey emotion, age, and gender. The most important features for determining these characteristics are facial landmarks, such as the shapes of the eyes, lips, nose, jawline, eyebrows, and wrinkles [3]. These characteristics vary subtly among individuals and differ between males and females. As one ages, facial muscles can also contract. Six basic human emotions have been classified: happy, sad, surprise, anger, fear, and disgust [4]. Although six basic emotions have been classified, Robert Plutchik's theory presents a different classification with eight emotions: fear, anger, sadness, joy, disgust, trust, anticipation, and surprise [5]. Meanwhile, Book Two of Aristotle's Rhetoric suggests nine types of emotions: anger, friendship, fear, shame, kindness, pity, indignation, envy, and love [6].

Based on previous studies, several issues have been identified in emotion, age, and gender recognition [7][8][9]. The efficacy of current methods for recognizing emotion, age range, and gender is still primarily evaluated using static images [9][10] Different methods lead to different outputs and results. Therefore, finding accurate and specific methods for these recognition tasks requires a significant amount of time for training. Humans are able to show different types of emotions, especially the six basic emotions: happy, sad, anger, fear, surprise, and disgust. Each of these emotions can be recognized from the facial expressions made by an individual based on their situation or mood. Unclear videos or images can negatively affect the recognition process and, consequently, the accuracy of the results [7][8][11].

Additionally, age recognition is the most difficult of the three tasks, due to the various age ranges that must be categorized. It is also the most difficult to perform physically [12] A human cannot recognize the exact age of an individual based on their expression alone. In contrast, emotion and gender can be recognized more easily than age due to distinct emotional characteristics and gender-specific facial features. However, this can be challenging, as both genders may show the same emotion in similar situations [11]. In this study, "recognition" refers specifically to the task of identifying a person's emotion in response to certain circumstances. Deep learning, using a new two-tier method with an artificial neural network (ANN) and a Local-deep neural network (LDNN) algorithm, has been used to perform emotion, age, and gender recognition in

both training and real-time modes [13]. This was accomplished using two different datasets: the AffectNet dataset and the UTKFace dataset. The images from these datasets are used to test the system and verify correct outputs before it is deployed for real-time recognition.

2. RELATED WORK

A review of previous studies on emotion, age, and gender recognition provides an overview of sources that can be explored when researching this topic. These studies have used a variety of algorithms, methods, and techniques to develop systems that can recognize emotion, age, and gender [14][15][29]. Moreover, this body of work helps to enhance the accuracy of emotion, age, and gender recognition using deep learning. In recent years, this topic has become popular in different fields, especially in security [16]. Emotion, age, and gender can be recognized from a single image with the help of accurate algorithms such as the Convolutional Neural Network (CNN) and the Local-Deep Neural Network (LDNN). For example, the EAGR System used a CNN algorithm with the NFC and BLAC methods to achieve an accuracy of 74.9% for emotion recognition, 69.15% for age recognition, and 91.75% for gender recognition [17] In contrast, the DAGER system increased the accuracy of emotion recognition to 93.2%, age recognition to 76.1%, and gender recognition to 91% by using only the DNN algorithm with the Sighthound method [18]. Hence, the algorithms and methods used for the three recognition categories affect their respective accuracies. Table 1 shows a summary of previous studies on emotion, age, and gender recognition:

Table 1: shows the summary of previous study of emotion, age and gender recognition

Authors	Algorithms	dataset	Results of emotions / expression recognition	Result of age recognition	Result of gender recognition
[17]	CNN	N/A	- Original method: 74.1% - Pre-processing of face detection: 76.9% -Pre-processing of normalization:74.9%	- Original method: 63.1% - Pre-processing of face detection: 74.5% -Pre-processing of normalization:69.1%	- Original method: 91.9% - Pre-processing of face detection: 93% -Pre-processing of normalization:91.7%
[2]	DNN	LFT-deep training and ChaLearn- data Pre-processing	93.2%	76.1%	91%
[19]	CNN	- Adience (age and gender recognition) - AfffectNet (emotion recognition)	67.65%	62.11	91.8%
[20]	Viola jones & LBP	Development of human–computer interaction (HCI) applications	86.2%	N/A	N/A
[21]	CNN & Deep network	MEGVII face classification	66%	N/A	N/A

3. Proposed Method

This study uses two approaches for recognizing a person's emotion, age, and gender: one using static images and another using real-time video. A general overview of the approach for facial emotion, age, and gender recognition is illustrated in Figure 1. First, a human face from a still image or a real-time video is provided as input to the system. The input image or video is then pre-processed to detect the face and remove background noise. This process uses methods such as filtering and data augmentation. After the face is processed from the input image, features are extracted to differentiate it from others. Then, the process continues with system testing to ensure it meets all requirements without any issues. The system is tested using images from two different datasets: the AffectNet dataset [16] and UTKFace dataset [22].

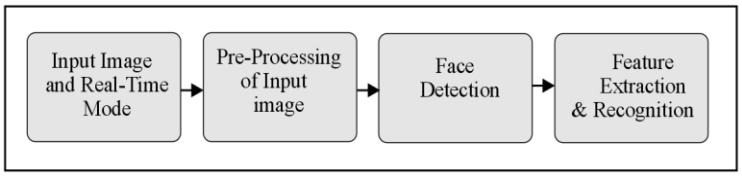

Figure 1: A general approach for emotion, age and gender recognition

The system will be tested with images from the dataset to ensure correct outputs before its deployment for real-time recognition.

3.1 workflow model of training

Before real-time recognition, a training phase is required to ensure the system can perform the required tasks. Figure 2 shows the workflow for the training model, from the initial process to the final result.

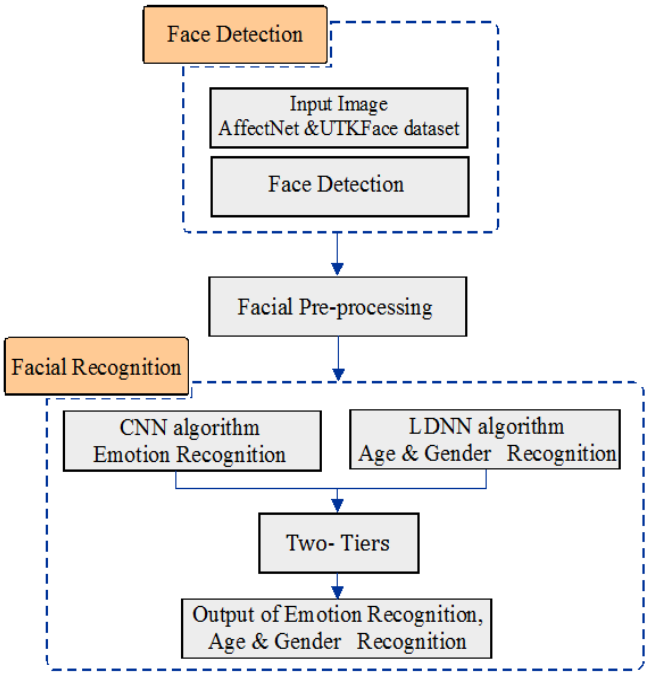

Figure 2: Workflow model of the system for training

Initially, input images are taken from open online datasets: the AffectNet dataset is used for emotion recognition, and the UTKFace dataset is used for age and gender recognition. Next, a face is detected in the input image to proceed with the face processing phase. If a face cannot be detected in an image, the system will not be able to produce an accurate recognition of emotion, age, and gender. Additionally, this will result in a processing error. Once a face is detected, the image proceeds to the facial recognition phase. In this phase, the face image is fed into two different algorithms depending on the recognition task. Emotion recognition uses a CNN algorithm to detect the emotion, while an LDNN algorithm classifies the age and gender from the input image. Both are deep learning algorithms that follow a similar process. Each algorithm has multiple layers capable of processing facial features to output the emotion, age, and gender recognition results from an input image. The combination of these algorithms leads to a two-tier architecture, where both are used to generate a single output for the system. Finally, for any image where a face is successfully detected, the system will output the resulting emotion, age, and gender of the person in the image.

After the training phase is complete, the system is able to recognize emotion, age, and gender, and can then proceed with real-time recognition. The workflow for real-time recognition is quite similar to that of the training model. However, an initial input image is not needed, as the system immediately detects any face it can capture in real-time. Once a face is detected, the system proceeds with facial pre-processing. In this phase, the detected face is processed in detail by analyzing facial features such as the eyebrows, eyes, nose, mouth, and jawline. The process involves scanning from the top of the face to the bottom. Each facial feature

is important and must be processed to ensure the accuracy of the emotion, age, and gender recognition. Then, similar to the training workflow, the detected face is fed into two different algorithms: the CNN algorithm and the LDNN algorithm. Next, the two algorithms are combined in the two-tier architecture to provide a single output that recognizes the emotion, age, and gender.

3.2 Two- Tiers Architecture

After that, the processed data is pushed to a two-tier architecture that separates the components of the flow into different locations [23]. In the two-tier architecture method, the user interface is located on the client's desktop application, while the application logic and database management services reside on a powerful server that serves many clients [24]. The first tier handles the input and user interface, while the second tier serves data and executes the application logic.The output results from the facial recognition phase. The face is recognized using two algorithms, both of which help to identify a person's specific emotion, age, and gender.

i. CNN Algorithm

A Convolutional Neural Network (CNN) is an algorithm with multiple layers capable of processing facial features to output an emotion recognition result from an input image, as shown in Figure 3.

The primary layers are known as convolutional layers. These layers perform a convolution operation on the input data. Within these layers, a filter performs the convolution, and a Rectified Linear Unit (ReLU) then performs an operation on the resulting elements. This feature extraction process generates feature maps.

Let f_k be the filter with a kernel size of $n \times m$ applied to input x. $n \times m$ is the number of input connections each CNN neuron has. The resulting output of the layer calculates as below equation [25]:

$$C(Xu,u) = \sum_{i=-\frac{n}{2}}^{\frac{n}{2}} \sum_{j=-\frac{m}{2}}^{\frac{m}{2}} f_k(i,j) xu - i, v - j \quad (1)$$

To generate a diverse input representation, multiple filters f_k with $k \in N$ can be employed. These filters are implemented by sharing weights among neighboring neurons. This weight-sharing approach offers a significant benefit: it reduces the number of independent weights that need to be trained compared to a conventional Multilayer Perceptron, as many weights are tied together.

Each feature map is then sampled in the pooling layer [26]. The pooling layer converts the two-dimensional arrays from the pooled feature map into a single, long, continuous, linear vector by flattening it. While there are several types of pooling, this study uses max pooling. Max pooling reduces the input's dimensionality by applying the maximum function over a given region. Let m be the size of the filter then the output calculates as follows:

$$M(x_i) = max\{x_i + k, i + l \mid |k| \leq \tfrac{m}{2}, |l| \leq \tfrac{m}{2} k, l \in N\} \quad (2)$$

The Rectified Linear Unit (ReLU) is an activation function used in neural networks. It calculates the output from a given input x, using the function below:

$$R(x) = \max(0, x) \quad (3)$$

This function helps prevent the vanishing gradient problem, as its derivative is constant for any positive input [27]. At the end of the process, the algorithm uses a series of fully-connected layers, which form a multilayer perceptron (MLP) network. These layers connect every neuron from the previous layer to every neuron in the current layer.For a given fully-connected layer, the input is x while the size will be k and the number of neurons in the fully connected layer will be l This structure results in a weight matrix $W_l \times k$ and σ is called as activation function where in this network σ is the identity function.

$$F(x) = \sigma(W * x) \quad (4)$$

Furthermore, there is an output layer that represents the predicted class for a given input image. The features gathered from the preceding layers provide a comprehensive representation of the original image, culminating in this final classification. The resulting class for the output vector x is shown below:

$$C(x) = \{i \mid \exists i \, \forall j \neq i : xj \leq xi\} \quad (5)$$

Thereafter, the $SoftMax$ function is use in the final output layer [28]. It calculates the probability for each class, and the sum of the probabilities for all classes equals one. From the final output, the class with the maximum probability is selected as the predicted output. The SoftMax function is defined by the equation below:

$$S(x)i = \frac{e^{xi}}{\sum_{i=1}^{N} e^{xi}} \quad (6)$$

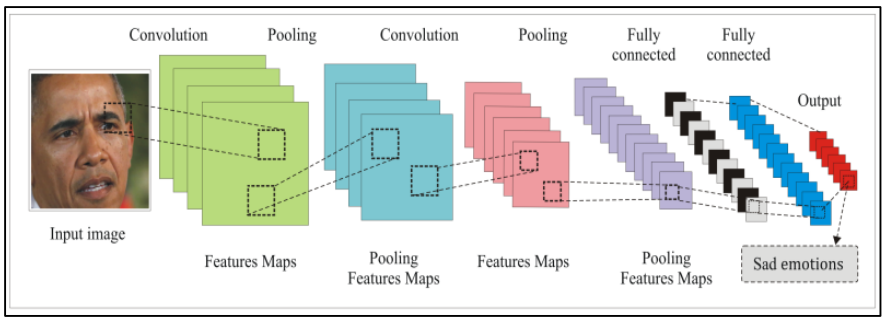

Figure 3: CNN workflow

ii. LDNN Algorithm

Local Deep Neural Networks (LDNNs) employ a distinct training strategy compared to Convolutional Neural Networks (CNNs). LDNNs extract small image patches, typically numbering in the hundreds, centered around significant regions. These patches are then used as input for age and gender recognition. Specifically, the eye and mouth areas are critical for age recognition, whereas gender recognition primarily focuses on the eye regions. This process is illustrated in Figure 4. Initially, an edge detection filter is applied to the images to identify edges. Subsequently, patches are extracted around these detected edges and fed into a deep neural network for training. During testing, the predictions from all patches derived from a single image are averaged to produce the final prediction for that image. This method of using filtered patches reduces the likelihood of overfitting by eliminating much of the redundant information. Consequently, a feed-forward neural network without dropout is utilized.

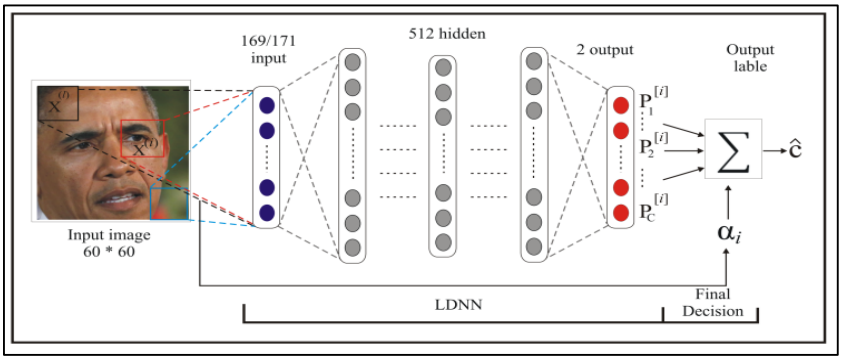

Figure 4: Training of LDNN

As previously mentioned, LDNN generates hundreds of patches for each image, which can be computationally expensive for training a neural network with thousands of images. To mitigate this cost, the nine-patch method has been proposed. This method reduces the number of patches generated to only nine per image. As illustrated in Figure 5, these nine patches are indexed from left to right and top to bottom. The

top-left patch is labeled as the first patch, and the bottom-right patch is labeled as the ninth patch. Furthermore, the height and width of these patches are set to half of the original height and width of the input images.

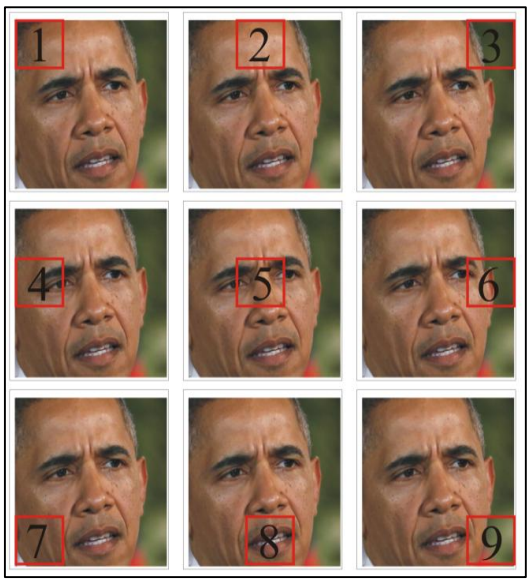

Figure 5: The nine patches method

4. Result and Discussion

The results are divided into two categories: training recognition and real-time recognition. For the training portion, the system used datasets as the source for input images. The images from the AffectNet dataset [18] and UTKFace dataset [19] that were fed into the program produced outputs for the emotion, age, and gender of a person from the still images. The system successfully recognized the six basic emotions (sad, fear, anger, surprise, disgust, and happy), along with different age ranges and both genders (male and female). The training phase successfully produced outputs for all three tasks, with emotions and genders being recognized accurately. Therefore, the objective of increasing the accuracy of emotion recognition was achieved, as the system was able to recognize all six basic emotions from the still images.

Figure 6: The results of emotion, age and gender from AffectNet datasets

The result is output immediately when a person's face is detected by the webcam. Figure 7 shows the real-time recognition system being run and tested. The system was tested on various people—such as lecturers, staff members, friends, and industry professionals—to recognize their emotion, age, and gender. It demonstrated high accuracy for recognizing emotion and gender. However, age recognition remains the most challenging task, likely due to subtle variations in facial features.

Figure 7: The results of emotion, age and gender in real – time mode

4. Conclusion

This paper proposes a new two-tier architecture that combines a Convolutional Neural Network (CNN) and a Local Deep Neural Network (LDNN). The results are demonstrated in both training and real-time recognition phases. The training phase was successful, as the system could detect all basic emotions, different age ranges, and both genders. This demonstrates the effectiveness of combining the CNN and LDNN algorithms for accurately detecting these three characteristics. However, certain circumstances negatively affected the performance during real-time recognition. Accuracy decreased due to factors like background lighting, noise, and ambiguous facial features, which made it difficult to determine a person's exact age. Some individuals appeared older than their actual age, while others appeared younger. This study's primary novelty is its specific two-tier system for recognizing six basic emotions, eight age groups, and two genders. The results revealed that with the CNN algorithm, emotion recognition accuracy reached 95.7%. Moreover, gender recognition presented little difficulty, as it can be differentiated by distinct facial features. Finally, age recognition proved to be the most difficult task because people often look younger or older than they are, which resulted in lower accuracy for this category. In summary, while the combined algorithms showed promising results, the process requires refinement. Further training is needed to increase the system's overall effectiveness, particularly in real-world conditions.

Acknowledgment

This real – time recognition is run and has been tested during the IREX 2022 event that organized by Faculty of Information Sciences and Engineering in Management and Science University, we would like to thanks and appreciate their supporting.

References

[1] Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M. A., & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(01), 73-79.

[2] Mohamed Naleer, H. M. (2020). Human face recognition to target commercial on digital display via gender.

[3] He, L., Zhang, P., Kong, J., Bai, H., Wang, Y., & Ding, X. (2024). Gender differences in individual emotion recognition in threatening situations: An eye-tracking study. *Current Psychology*, 43(29), 24595-24607.

[4] Ferreira, B. L. C., de Morais Fabrício, D., & Chagas, M. H. N. (2021). Are facial emotion recognition tasks adequate for assessing social cognition in older people? A review of the literature. *Archives of gerontology and geriatrics*, 92, 104277.

[5] Abdul-Mageed, M., & Ungar, L. (2017). Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (volume 1: Long papers) (pp. 718-728).

[6] Latha, C. P., & Priya, M. (2016). A review on deep learning algorithms for speech and facial emotion recognition. *APTIKOM Journal on Computer Science and Information Technologies*, 1(3), 92-108.

[7] Baffour, P. A., Nunoo-Mensah, H., Keelson, E., & Kommey, B. (2022). A survey on deep learning algorithms in facial Emotion Detection and Recognition. *Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, 7(1), 24-32.

[8] T.-T. Lu, S.-C. Yeh, C.-H. Wang, and M.-R. Wei, “Cost-effective real-time recognition for human emotion-age-gender using deep learning with normalized facial cropping preprocess,” *Multimed Tools Appl*, vol. 80, no. 13, pp. 19845–19866, 2021.
[9] Kumar, R., Corvisieri, G., Fici, T. F., Hussain, S. I., Tegolo, D., & Valenti, C. (2025). Transfer Learning for Facial Expression Recognition. *Information*, 16(4), 320.
[10] Chen, L., Wu, M., Pedrycz, W., & Hirota, K. (2020). Emotion-Age-Gender-Nationality Based Intention Understanding Using Two-Layer Fuzzy Support Vector Regression. In *Emotion Recognition and Understanding for Emotional Human-Robot Interaction Systems* (pp. 183-214). Cham: Springer International Publishing.
[11] Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information* fusion, 59, 103-126.
[12] Abushahma, R. I. H., Ali, M. A., Al-Sanjary, O. I., & Tahir, N. M. (2019, December). Region-based convolutional neural network as object detection in images. In *2019 IEEE 7th Conference on Systems, Process and Control (ICSPC)* (pp. 264-268). IEEE.
[13] Chindiyababy, U., Kakkar, P., Vedula, J., Yunus, J., Umidbek, A., & Sharma, S. (2025). Deep Learning-Based Facial Emotion Recognition for Advanced Human-Computer Interaction. In *2025 First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies* (CE2CT) (pp. 1247-1252). IEEE.
[14] Aiswarya, P., & Mangalraj, P. (2020). Emotion recognition by inclusion of age and gender parameters with a novel hierarchical approach using deep learning. In *2020 Advanced Communication Technologies and Signal Processing (ACTS)* (pp. 1-6). IEEE.
[15] Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3), 1195-1215.
[16] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1), 18-31.
[17] Mondal, J., & Deshpande, A. (2014). Eagr: Supporting continuous ego-centric aggregate queries over large dynamic graphs. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of data* (pp. 1335-1346).
[18] C.-M. Tsai and F. Y. Shih, “An efficient detection and recognition system for multiple motorcycle license plates based on decision tree,” *Intern J Pattern Recognit Artif Intell*, vol. 36, no. 05, p. 2250022, 2022.
[19] Tsai, C. M., & Shih, F. Y. (2022). An efficient detection and recognition system for multiple motorcycle license plates based on decision tree. *International Journal of Pattern Recognition and Artificial Intelligence*, 36(05), 2250022.
[20] Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116, 56-76
[21] Masi, I., Trần, A. T., Hassner, T., Leksut, J. T., & Medioni, G. (2016). Do we really need to collect millions of faces for effective face recognition? In *European conference on computer vision* (pp. 579-596). Cham: Springer International Publishing.
[22] Chandaliya, P. K., Kumar, V., Harjani, M., & Nain, N. (2019). Scdae: Ethnicity and gender alteration on clf and utkface dataset. In *International Conference on Computer Vision and Image Processing* (pp. 294-306). Singapore: Springer Singapore..
[23] Haq, H. B. U., Akram, W., Irshad, M. N., Kosar, A., & Abid, M. (2024). Enhanced real-time facial expression recognition using deep learning. *Acadlore Trans. Mach. Learn*, 3(1), 24-35.
[24] AlBarki, M., Alardawi, A., Aboshgifa, A., & Belhaj, N. (2024). Classes Scheduler sing Genetic Algorithm. *Science and Technology*, 12(01).
[25] Dhruv, P., & Naskar, S. (2020). Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): A review. *Machine learning and information processing: proceedings of ICMLIP 2019*, 367-381.
[26] Jie, H. J., & Wanda, P. (2020). RunPool: A dynamic pooling layer for convolution neural network. *International Journal of Computational Intelligence Systems*, 13(1), 66-76.
[27] Kessler, T., Dorian, G., & Mack, J. H. (2017). Application of a rectified linear unit (ReLU) based artificial neural network to cetane number predictions. In *Internal combustion engine division fall technical conference* (Vol. 58318, p. V001T02A006). American Society of Mechanical Engineers
[28] Jiang, M., Liang, Y., Feng, X., Fan, X., Pei, Z., Xue, Y., & Guan, R. (2018). Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*, 29(1), 61-70.
[29] Haq, H. B. U., Akram, W., Irshad, M. N., Kosar, A., & Abid, M. (2024). Enhanced real-time facial expression recognition using deep learning. *Acadlore Trans. Mach. Learn*, 3(1), 24-35.