Mapana – Journal of Sciences
2023, Vol. 22, Special Issue 2, *297-311*
ISSN 0975-3303 | https://doi.org/10.12723/mjs.sp2.17

An Intelligent Facial Recognition System using Stacked Auto Encoder with Convolutional Neural Network (CNN) Approach

N. Mahendiran*

Abstract

The act of identifying an emotional feeling or state is described as facial expression. It is one of the effective techniques for interperson communication. They serve as indications that regulate interactions with those around. As a result, they are crucial in creating effective relationships. The goal of the facial expression recognition system is to identify the expressions by evaluating the changes in facial characteristics and extracting features from facial images. This system is essential for enhancing computer-human interaction. The majority of facial emotion recognition research mainly relies on a reference face model and well-known facial landmarks. Due to the intricacy of the face musculature, finding the most noticeable facial landmarks can be difficult and requires physical intervention for improved accuracy. Model based approaches need to establish a reference model and complex functions for mapping which takes intense computation time. So, this research work provides a new dimension to deal with the above issues by proposing a Stacked Auto-Encoder with Convolutional Neural Network based approach that does not rely on the landmarks or a reference model. The proposed approach outperforms the existing techniques.

Keyword: Facial recognition, geometric feature, deep learning, auto-encoder, neural network, and classification.

* Bharathiar University, Tamil Nadu, India; Email: nmahendran.snr@gmail.com

1. Introduction

In recent years, a large establishment in artificial intelligence has advanced the intensive study of automatic recognition systems using facial expressions. The most ancillary nonverbal way for humans to communicate with each other is through facial expressions and it was initially suggested [1]. To shore up facial expressions, suggested six basic expressions such as happy, sad, anger, surprise, fear, and disgust. The happy expression is indicated by a smile on a human and it signifies a curved shape eye, lip corners pulled up and back, raised cheek, and wrinkles created in the outer nose and beside the eyes. The sad expression is normally expressed as rising skewed eyebrows, corner of lips pulled down and inner corner eyebrows angled upwards. The anger expression is explicated with squeezed eyebrows, eyebrows pulled together, lips pressed together, and slender and stretched eyelids. The disgust expression is indicated with pull down eyebrows, narrowed eyes, cheeks raised, and a creased nose [2, 3].

The surprise expression is expressed with widened eyes, mouth open, and horizontal forehead wrinkles [4]. The expression of fear is explicated with opened mouth, lips stretched, and skewed eyebrows. Eminently majority of FER systems were prompted to recognize six basic emotional expressions and certain more emotions such as contempt, envy, pain, and drowsiness were attempted by a few FER systems [5]. Furthermore, some FER systems associate expressions with spontaneous and pose-based expressions. The facial expression that a human express certainly on the face during their daily routine life such as delivering dialogues, watching movies, etc. is termed a spontaneous expression [6, 7].

The pretended expressions by humans generally attained in a laboratory environment are termed pose-based expressions [8]. The appearance, timing, intensity, and dynamics of pose-based expressions are superlative in comparison with spontaneous expression. Furthermore, the next two categorizations of facial expressions, micro-expressions, and macro-expressions, have received more embodiment in recent research works [9]. Micro-expressions are involuntarily and unintended facial expressions, which last for less than half a second and invoke a person's true

emotion. The macro-expressions are usually the normal expression of humans which can last for half second to 4 seconds [10].

Recently, emotion recognition has been turned to be an active research topic in affect computing. For human beings, emotion enacts a substantial aspect of their daily life. Emotion in humans is evinced by the mental state, behavioural acts and physiological changes and hence it should be monitored [11]. Also, the activities and hobbies of individuals are very well assured by the emotion they incurred. Facial expression tends to be an effective nonverbal communication tool that is likely to specify emotional behaviour, mind intention, and sentiments. Persons can easily communicate information using facial expressions even without speaking [12]. The deformation of face parts and its structural affinity or changes in the appearance of the face are implicated in the analysis of automatic facial emotion recognition. Moreover, facial expressions signify the transition of facial appearance in accordance with social communication, inner emotional state, or intentions [13].

The Computer Vision and Machine Learning are branches of Artificial Intelligence that describes the method of training the systems with intelligent algorithms and protocols [14]. The algorithms and protocols define the set of rules to be followed to work as if the humans work in certain situations such as identification of humans, recognition of objects, after recognition of objects, the systems may work as such humans work in certain environments that require intelligence. Thus, the research has given its attempt to identify the facial similarities among humans in both controlled and un-controlled conditions as a part of fulfilling the gaps identified in Section 2. Face detection and Face Recognition is another method of machine learning that requires algorithms and protocols to detect and recognize humans. Such detection and recognition of faces involves employing suitable algorithms [15]. Mobile devices or Hand-held devices are used for communication as well as biometrics as a part of security purposes. The biometrics requires the detection and recognition of faces of owner of the hand-held devices to unlock [16].

In order to carry-out these biometric activities for the purpose of securing a device or any other authentication purpose requires the employability of algorithms of Machine Learning to make the

system Artificially Intelligent [17]. The artificially intelligent system is one that works more efficiently, if we have deployed an efficient algorithm in devices. Hence, research focussed attention to develop an efficient algorithm to determine the facial similarities among relationships of images. The facial kinship or Facial similarity is a task achieved by analysing the facial features present in an image. Initially, the biometric device required to achieve face detection, face recognition and facial similarity verification to accomplish the process of biometric security [18].

Facial Image Similarity is very essential in the study of biometrics, as it requires processing of facial features. The facial features may include eyes, nose, mouth, chin, cheek, and jaws as a part of facial features verification. If the facial features of actual person are correctly determined the system allows the user to unlock the device in biometric enabled devices. Similarly, the same facial identification shall be used in unlocking the doors in officers instead of barcode scanners in identity cards.

The research article is organized as follows: the overview of facial recognition is given in Section 1, the literature review is detailed in Section 2, the proposed methodology is given in Section 3, result and discussion is given in Section 4, and the article is concluded in Section 5.

2. Related Works

In today's scientific research, facial expression recognition has become a huge topic of interest. Facial emotions are the most expressive way to convey human emotions. Facial emotion recognition is a biometric technique for identifying human emotion based on a facial image. FER systems find major applications in bio medics, data analytics, human-computer interaction, assisting behaviour understanding research, and deceit detection. The steps occupied in facial emotion recognition are image acquisition, pre-processing, face localization and detection, feature extraction, and classification. Once captured, the facial images are scaled and rotated to ensure that the size and orientation of the face are suitable for subsequent processes. The following stage is featuring extraction, in which the extracted features are used to generate a numerical map of each face under investigation. There are certain complications in

face images such as poor contrast, poor pose, occlusion problem, processing time, etc. Machine learning and optimization methods can be used to solve these issues. This chapter discusses the different methods and strategies used by different researchers for facial expression recognition [20].

Ko et al. (2018) proposed a hybrid approach was proposed in which a Convolutional Neural Network (CNN) that represents spatial features is combined with Long Short-Term Memory (LSTM) to represent temporal features of consecutive frames. LSTM possesses a chainlike structure that is designed to solve the dependency in long-term using short term memory. LSTM supports both fixed and variable length input and output. Moreover, they support straightforward end-to-end fine tuning when they are integrated with other models such as CNN [19].

Two different features namely geometric and appearance features are not likely to be related statistically as the nonlinearities in the function mapping are of different order. So, the simple conventional feature fusion approaches might perform worse because of the uneven representation of different domains of features. So, Majumder et al. (2018) proposed a fusion framework using auto encoders that learns correlation to represent features better. Geometric features were represented using eye projection ratio and directional displacement information in key facial regions. Appearance based features were represented using Regional LBP. The two features were fused together using auto encoders. Then, an improved Self Organizing Map based classifier was utilized to yield better results. Here, soft-threshold logic was used to avoid misleading of prediction [21].

A Hybrid Neural Network (HNN) that combined the features of unsupervised learning along with artificial Neural Network was proposed by Han Y et al. (2021) to enhance the network inference capability and interpretability. This was used for enhancing the social emotion classification capability. In addition to HNN, a computational tool of Latent Semantic Machines and Transfer Learning Function was developed to extract semantic features that disambiguate different context of emotions. A sparse encoding method that filtered the noisy images was also introduced to regularize the learning of semantic features. Thus, they empowered

Neural Networks with semantics to enhance the social emotion capability [22].

In certain datasets, the relationships between the characteristics are complicated. Therefore, using an existing technique and single Autoencoder is insufficient. It is possible that a single autoencoder will be unable to decrease the input features' dimension. Therefore, the research employs stacked autoencoders and neural networks for such application situations.

3. Proposed Methodology

Deep learning is the process in which different layers present in a network used for learning different features. These layers finally connected to the last layer from which the final output is achieved. So basically, this research utilized this layered architecture to extract the features without utilizing its final fully connected layer. Autoencoders are artificial neural networks that enable the application of ' blank ' data. Auto-encoder attempts to learn a conversion into a compact and distributed illustration for a particular set of inputs. Therefore, Autoencoders may be seen as dimensional (or failure in compression). One automated coder is a layer of information that corresponds to the raw data display, along with a hidden layer of encoding. Figure 1 depicts a stacked autoencoder that readily connects the outputs of the individual layer to the following layer of input. The encoding is usually conducted in an autoencoder, which minimizes the dissimilarity factor (reconstruction error) between the data and that of the hidden layer (using inverse weights).

One of the most valuable benefits of stacked automatic encoders: ease of use layer by layer, which enables them to learn from unchecked input encoding.

Assume that a matrix which is 2-D in size I x J and the size of $\hat{I} \times \hat{J}$ for convolutional layer translate into a two-dimensional input room, with a stride $\hat{I}\&\tilde{J}$. This layer$\tilde{}$ brings into being the size of the output $\frac{I-\hat{I}}{\tilde{I}} \times \frac{J-\hat{J}}{\tilde{J}}$. The neurons layers calculate their activations based on equation 1.

$$yi,j = f\left(\sum_{i=0}^{\hat{I}-1}\sum_{j=0}^{\hat{J}-1} K_{\hat{\imath},\hat{\jmath}} x_{i+\hat{\imath},j+\hat{\jmath}}\right)\forall i \in [1,\frac{I-\hat{I}}{\hat{I}}), j \in 1,\frac{J-\hat{J}}{\hat{J}})$$

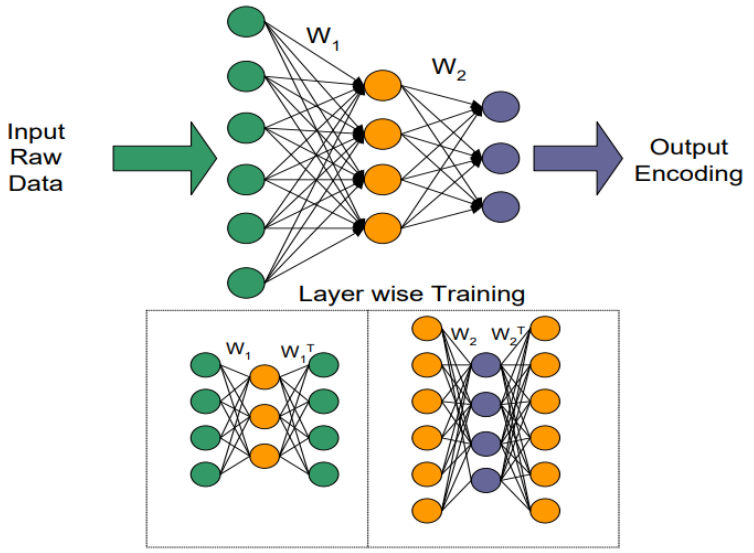

Figure 1. Stacked Auto Encoder

Where the i^{th} row and j^{th} column in the input and output space are denoted by $x_{i,j}$ and $y_{i,j}$ the primary feature of convolutionary layers is the combination of the filter weights at different points throughout the input.

The manufacture of a convolutionary layer conclusively indicates the continuation of functions defined in the input space in an explicit place. If the place of this function in the entry is translated, the layer activation rate will then always convert proportionally. The goal is to apply force to small input irregularities and give a degree of invariance to the filters. The bundling process usually involves the reading of a 2-D input patch and the creation of a single output calculated using the input patch feature. This spreads across other fields of information, preserving the entire data region. It is important to note that for a pooling process, no weights are necessary; this simply acts as a sub-sample step. Max pooling is a typical kind of pooling in which output compares with its input number. This would be the function calculated by using a given 2-D input patch X to define the output of a max-pooling procedure:

$$y = \max(X)$$

It cannot be regarded as a place on the input room but rather as a function in the field of the input area. Standardization of the local contrast is a means of enforcing rivalry between neighboring neuron activations. A nearby neuronal patch minimizes stimulation. Subtractive and divisive criteria combine this standardization: divide the outcomes by standard values deviations and remove the mean of the local patch from each value. The following words can be described:

Contrast normalization can be carried out by a local input image of Size I x J

$$y_{i,j} = \frac{x_{ij} - \frac{1}{IJ}\sum_{i\in I, j\in J} x_{i,j}}{\sqrt{\sum_{i\in I, j\in J}\left(x_{i,j} - \frac{1}{IJ}\sum_{i\in I, j\in J} x_{i,j}\right)^2}} \; \forall i \in I, j \in J$$

Where $x_{i,j}$ and $y_{i,j}$ corresponds to the input and output situated in the input and output area at the i^{th} row and j^{th} column.

This type of training consists of not only changing or replacing the final layers; preferably, it also includes the fine-tuning of previous layers and retraining them. Deep networks are those networks that are highly configurable, and that can be done by using various hyperparameters. The initial layers of the deep network are generally used to capture generic features while the later ones focus more on the specific task in hand. Following figure 2 shows the face recognition problem in which the initial layers are to learn very generic features, and the subsequent layers are learning task-specific features. Using the insight, individual weights can be fixed and used for retraining, and remaining can be finetuned as per the requirements. In such models, the knowledge has been utilized, and the state of the network is utilized as a starting point for retraining purposes. So, better performance can be achieved with such models in less training time.

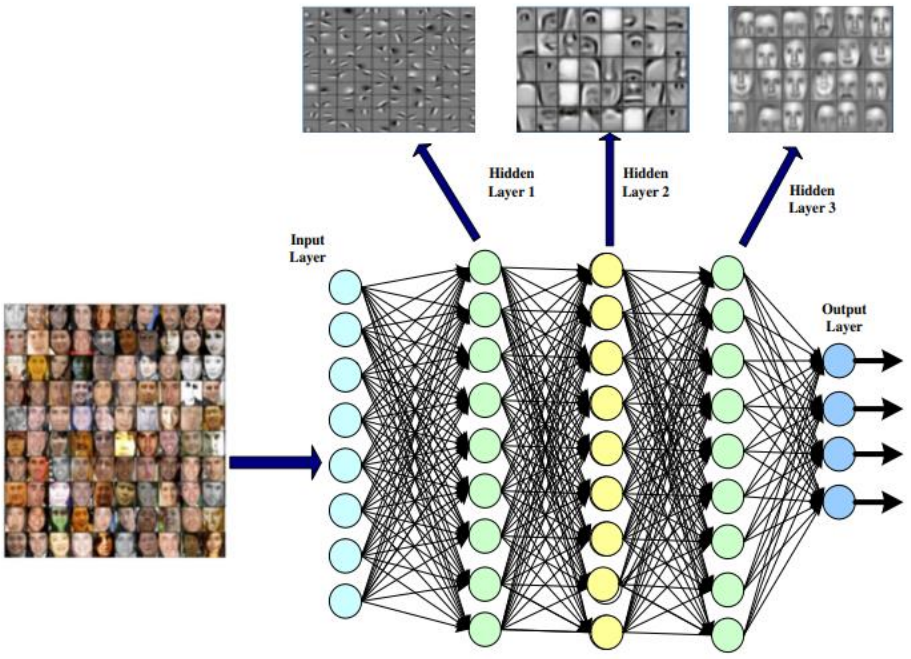

Figure 2. Representation of Stacked Auto Encoder with CNN

4. Result and Discussion

This section explains about the process of feature selection and discussion about its outcome. The numerical outcome and graphical illustrations of classification process is detailed along with comparative analysis. The research work is implemented using the MATLAB and utilises publicly available dataset.

4.1 Feature Selection

Geometric features represent information about the variations in position, shape and distance of prominent facial components like eyes, eyebrows, mouth, nose, etc. These features need more accurate method for facial feature detection as they are noise sensitive, which is difficult to use in many real-world situations.

Appearance based features represent changes in appearance of the face such as wrinkles, furrow etc. Here, image filters are applied to

extract features from the whole facial image or from the specific regions of a facial image to capture changes in face appearance. These features are less dependent on initialization and can represent the micro patterns in the texture of the skin, which is an important cue for FER. But it is difficult to generalize the appearance features across individuals.

It uses both the geometric and appearance approaches for extracting the features. Thus, it combines the advantage of both the approaches resulting in a better accuracy. Human beings, in addition to the identification of appearance features, can also reason out using facial geometric features like widening of lips and eyebrows, sharpness of nose, wrinkles, openness of eyes etc.

4.2 Classification

Using the given set of data, process of predicting the class for a data is said to be termed as classification. Classes are also called as categories, labels or targets. Classifiers can be expressed as a function that maps input data representing features into a discrete output variable, representing class. They use a subset of input data as training samples for understanding how the input maps to the class. Once it understands, the mapping can be used to predict the class of unseen, testing sample. It is said to be in the category of supervised learning where the target class is provided with the given set of data.

The types of learners in classification can be of two types: lazy and eager. Lazy learners wait with training data until testing data appear. Eager learners construct a model using training data before testing data appears. Usually they take long training time, but less predicting time. Lot of classifiers are available and choice of the classifier depends on the application and nature of data e.g. linearly separable property.

To verify the applicability of the classifier, many methods are available. The common methods used for this are hold out and cross validation. In holdout method, given set of data will be divided according to the ratio of 80:20 (training: testing). Training samples will be used to build a model whereas the testing samples will be used to test the predicting power of the model. In cross validation method, the given set of data will be partitioned into 'k' mutually exclusive subsets of equal size. Out of this, one subset will be used

for testing while other subsets will be used for training. This will be repeated through all the 'k' subsets. The classification accuracy for the proposed and existing technique is given in Table 1.

Table 1. Comparison of Accuracy

Emotions	Classification Approach	Cropped Face	Whole Face
Anger	SAE-CNN	99.17	100.00
	CNN	100.00	97.50
	HNN	98.33	98.33
	LSTM	79.16	86.66
Disgust	SAE-CNN	100.00	100.00
	CNN	99.17	98.33
	HNN	96.66	92.50
	LSTM	87.50	83.33
Happy	SAE-CNN	99.17	100.00
	CNN	100.00	98.34
	HNN	99.16	99.16
	LSTM	93.33	81.66
Sad	SAE-CNN	100.00	100.00
	CNN	96.67	91.67
	HNN	99.16	92.50
	LSTM	80.00	75.00
Surprise	SAE-CNN	100.00	100.00
	CNN	99.17	100.00
	HNN	97.50	91.60
	LSTM	93.33	80.83
	DT-CWT	95.84	96.67

Emotions	Classification Approach	Cropped Face	Whole Face
Fear			
	CNN	96.67	88.34
	HNN	95.00	78.33
	LSTM	79.16	75.83

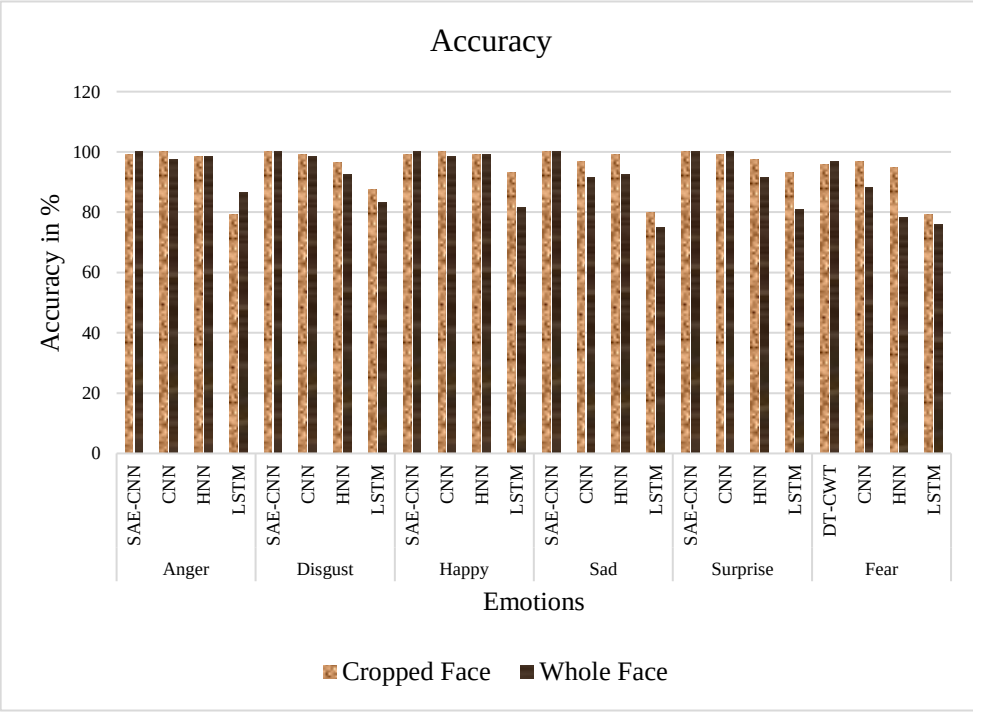

Figure 3. Comparison of Accuracy

5. Conclusion

Deep learning has emerged as a significant method with several applications. Here, the idea of deep learning is employed to identify face emotions. The problems that occur when implementing the suggested approach were the focus of this research. The focus of the ongoing study is on 3D FER, which uses a deep learning architecture and inputs highlighted photographs. The experimental results

demonstrate that the proposed work outperforms earlier state-of-the-art approaches. The proposed Convolutional Neural Network (CNN)-based technique to learn the attributes does not need raw facial pictures. In order to extract features from the data, CNN uses a stacked auto-encoder. This study shows that facial expression identification can be successfully accomplished utilising a CNN-based approach with input from highlighted photographs, and the learned features are better suited for categorization.

References

[1]. Singhal, P., Srivastava, P. K., Tiwari, A. K., & Shukla, R. K. (2022). A Survey: Approaches to facial detection and recognition with machine learning techniques. In *Proceedings of Second Doctoral Symposium on Computational Intelligence* (pp. 103-125). Springer, Singapore.

[2]. Onyema, E. M., Shukla, P. K., Dalal, S., Mathur, M. N., Zakariah, M., & Tiwari, B. (2021). Enhancement of patient facial recognition through deep learning algorithm: ConvNet. *Journal of Healthcare Engineering*, *2021*.

[3]. Fuad, M. T. H., Fime, A. A., Sikder, D., Iftee, M. A. R., Rabbi Rabbi, J., Al-Rakhami, M. S., ... & Islam, M. N. (2021). Recent advances in deep learning techniques for face recognition. *IEEE Access*, *9*, 99112-99142.

[4]. Hassan, R. J., & Abdulazeez, A. M. (2021). Deep learning convolutional neural network for face recognition: A review. *International Journal of Science and Business*, *5*(2), 114-127.

[5]. Shamrat, F. J. M., Al Jubair, M., Billah, M. M., Chakraborty, S., Alauddin, M., & Ranjan, R. (2021, June). A Deep Learning Approach for Face Detection using Max Pooling. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 760-764). IEEE.

[6]. 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*(02), 52-58.

[7]. Raju, K., Chinna Rao, B., Saikumar, K., & Lakshman Pratap, N. (2022). An Optimal Hybrid Solution to Local and Global Facial Recognition Through Machine Learning. In *A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems* (pp. 203-226). Springer, Cham.

[8]. Setta, S., Sinha, S., Mishra, M., & Choudhury, P. (2022). Real-time facial recognition using SURF-FAST. In *Data Management, Analytics and Innovation* (pp. 505-522). Springer, Singapore.

[9]. Coe, J., & Atay, M. (2021). Evaluating impact of race in facial recognition across machine learning and deep learning algorithms. *Computers*, *10*(9), 113.

[10]. Alburaiki, M. S. M., Johar, G. M., Helmi, R. A. A., & Alkawaz, M. H. (2021, August). Mobile based attendance system: face recognition and location detection using machine learning. In *2021 IEEE 12th Control and System Graduate Research Colloquium (ICSGRC)* (pp. 177-182). IEEE.

[11]. Lv, X., Su, M., & Wang, Z. (2021). Application of face recognition method under deep learning algorithm in embedded systems. *Microprocessors and Microsystems*, 104034.

[12]. Ibrahim, B. R., Khalifa, F. M., Zeebaree, S. R., Othman, N. A., Alkhayyat, A., Zebari, R. R., & Sadeeq, M. A. (2021, April). Embedded system for eye blink detection using machine learning technique. In *2021 1st Babylon International Conference on Information Technology and Science (BICITS)* (pp. 58-62). IEEE.

[13]. Hebbale, S., & Vani, V. (2022). Real time COVID-19 facemask detection using deep learning. *learning*, *6*(S4), 1446-1462.

[14]. Zhang, L., Sun, L., Yu, L., Dong, X., Chen, J., Cai, W., ... & Ning, X. (2021). ARFace: attention-aware and regularization for face recognition with reinforcement learning. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *4*(1), 30-42.

[15]. Joseph, L. L., Shrivastava, P., Kaushik, A., Bangare, S. L., Naveen, A., Raj, K. B., & Gulati, K. (2021). Methods to

identify facial detection in deep learning through the use of real-time training datasets management. *EFFLATOUNIA-Multidisciplinary Journal*, *5*(2), 1298-1311.

[16]. Akter, T., Ali, M. H., Khan, M., Satu, M., Uddin, M., Alyami, S. A., ... & Moni, M. A. (2021). Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage. *Brain Sciences*, *11*(6), 734.

[17]. Geetha, M., Latha, R. S., Nivetha, S. K., Hariprasath, S., Gowtham, S., & Deepak, C. S. (2021, January). Design of face detection and recognition system to monitor students during online examinations using Machine Learning algorithms. In *2021 international conference on computer communication and informatics (ICCCI)* (pp. 1-4). IEEE.

[18]. Zofishan, M., Islam, K. A., & Ghazal, F. (2021). MACHINE LEARNING BASED CLOUD MUSIC APPLICATION WITH FACIAL RECOGNITION USING ANDROID STUDIO (MUSYNC). *American International Journal of Sciences and Engineering Research*, *4*(1), 36-52.

[19]. Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, *18*(2), 401.

[20]. Wu, B. F., & Lin, C. H. (2018). Adaptive feature mapping for customizing deep learning based facial expression recognition model. *IEEE access*, *6*, 12451-12461.

[21]. Majumder, A., Behera, L., & Subramanian, V. K. (2016). Automatic facial expression recognition system using deep network-based data fusion. *IEEE transactions on cybernetics*, *48*(1), 103-114.

[22]. Han, Y., Wang, X., & Lu, Z. (2021). Research on facial expression recognition based on Multimodal data fusion and neural network. *arXiv preprint arXiv:2109.12724*.