2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS) | 979-8-3503-3583-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICISCOIS56541.2023.10100563

Emotion Recognition using Autoencoders: A Systematic Review

M.Mohana
Department of Computer Science
Centre for Machine Learning and Intelligence (CMLI)
Avinashilingam University
Coimbatore, India.
mohana_cs@avinuty.ac.in

Dr.P.Subashini
Department of Computer Science
Centre for Machine Learning and Intelligence (CMLI)
Avinashilingam University
Coimbatore, India.
subashini_cs@avinuty.ac.in

Abstract – During recent decades, facial expression recognition is a hot area of research in deep learning and computer vision. However, numerous research has been done on emotion recognition through facial expression using deep neural networks and achieved remarkable results on image datasets. Moreover, convolutional neural networks (CNN) often require a large number of layers when extracting useful information from facial images, thus increasing the network complexity and training time. For overcoming the challenges in CNN, researchers have been employing autoencoders to recognize facial emotions. This paper reviews all available work on autoencoders using facial expression recognition, and variations in autoencoders. For this analysis, a literature review paper has been collected from IEEE, Scopus, and Web of Science databases. Furthermore, publicly available facial expression recognition benchmark datasets are discussed. In addition, this paper discusses how unsupervised autoencoder has been applied in classification problems. Furthermore, this comprehensive review will be helpful for young researchers in FER and provide an overview of autoencoder in facial emotion recognition using facial expressions.

Keywords: Facial Emotion Recognition, Autoencoder, Dimensionality Reduction, ANN, CNN

I. INTRODUCTION

Facial Emotion Recognition (FER) is a method for conveying human emotions. For several decades, it has been a prominent research field in computer vision and pattern recognition [24]. FER has been applied in many applications but is not limited to human mental state detection, applied in healthcare to recognize patient emotional instability, personalized learning, and monitor candidate stress levels during interviews. Nevertheless, facial emotion recognition still has some challenges such as illumination, pose variation, scale variation, and occlusion [4]. It is very easy for a human to identify another human's emotion on the face, but machines are unable to do so. FER consists of four stages: pre-processing, face detection, feature extraction, and classification. Fig.1. shows the whole process of the FER system. The face detection technique is used in this initial stage to find a face in the image, and the detected face is cropped and resized. Viola-jones algorithm [3] is one of the famous face detection algorithms utilized by many studies because of this easiness. It is divided into four sections Haar features, Integral images, AdaBoost classifier, and cascade classifier. Haar features are used to extract the features using line, horizontal lines, and four rectangle kernels. Integral images speed up the feature calculation process. AdaBoost classifier builds a strong classier based on available features. Finally, the cascade classifier discards the non-face parts in images. This whole process repeats while detecting faces in images. After detection, face images are to be resized and normalized for the next stage of the process. In the second step, feature points have been extracted from detected faces. Finally, based on the retrieved and trained features, the classifier model is employed to classify the emotions.

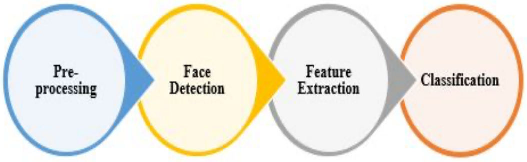

Fig. 1. Conventional FER approach

Many applications of machine learning algorithms rely on handcrafted extracted features from raw data. After the good features are extracted from the images, machine learning performs well on this data. This process of extracting features from images is called feature engineering. In FER, various feature-extracted techniques are commonly applied for extracting features [6] (e.g., HOG, LBP, SIFT). Sometimes it is hard to extract some minute features of facial expression like the micro expression on facial muscles. Deep learning methods such as deep neural networks (DNN), convolution neural networks (CNN), and recurrent neural networks (RNN) have minimized this issue [38]. This type of algorithm has performed well on the image recognition task because of its strong automatic feature extraction power. Nevertheless, due to the complex nature of facial expression images, CNN may necessitate a higher number of layers to retrieve certain relevant details from the data-representing face images. This increases the network depth, model complexity, and time complexity.

An important aspect of principal component analysis (PCA) is that it reduces the dimensionality of data [5]. It can be used for the linear transformation of high-dimensional code space to low-dimensional code space. Neural networks, on the other hand, feature at least one hidden layer that may be utilised to map the non-learner input layer to the output layer. But the conventional neural network is unable to compress the features to the same extent as a PCA. Autoencoder is an unsupervised artificial feedforward neural network with more than one hidden layer [1][37]. These networks are used to reconstruct the input as output at low dimensional. As a result, the size of the output layer is the same as the size of the input layer. Autoencoders are trained in a supervised manner using gradient decent backpropagation. As a result, the hidden layer is smaller than the input layer, and the dimensionality of the input layer is lowered and stored in the latent space's hidden layer. The output is reconstructed again from latent space into the original data. When using more than one hidden layer in an autoencoder neural network, the most relevant features are extracted and stored in a smaller code space. This can be given better results when building a classifier. In recent years, many researchers have utilized autoencoders in image classification problems [32].

The purpose of this review paper:

- To analyze the autoencoder and its variations, and
- To give a comprehensive review of unsupervised autoencoders in FER.

Fig.2. shows the number of previously published works on Autoencoder using FER datasets. This paper is structured as follows: Section 2 describes the basic architecture of the autoencoder and its variants. Section 3 describes available work on autoencoder for emotion recognition. Section 4 describes some commonly used facial expression datasets. Finally, Section 5 concludes this paper analysis review.

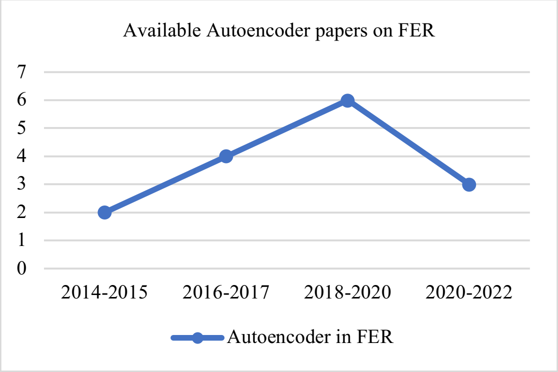

Fig. 2. Autoencoders in FER

II. ARCHITECTURE OF AUTOENCODER

Autoencoder was first designed in the 1980s and has played a significant role in the development of neural networks [1]. This was developed mainly for non-linear dimensionality reduction and the application of linear principal component analysis (PCA). The basic architecture of the Autoencoder is shown in fig. 3. In the earlier stage, autoencoders were used in pre-trained neural networks along with restricted Boltzmann machines that today are called the unsupervised deep learning algorithm. This pre-trained, neural network has assigned weights for each neuron learned by the autoencoder instead of random values. The dimension of the input and output vector is always the same as the autoencoder [2].

An autoencoder is a kind of artificial neural network that is capable of automatically learning key features from unsupervised learning. These are aimed to integrate input data into an internal latent representation and reconstruct output data that is comparable to the data input. The primary goal of learning is to provide an "informative" representation of the input data that may be utilised for a variety of applications such as clustering. Through unsupervised learning, the autoencoder finds useful structures from the data by disentangling sources of input data. When developing classifiers and prediction models, the resultant code space may be leveraged to obtain important information. As a result, when adequate data is used to train the autoencoder, a latent representation removes the requirement for substantial feature engineering.

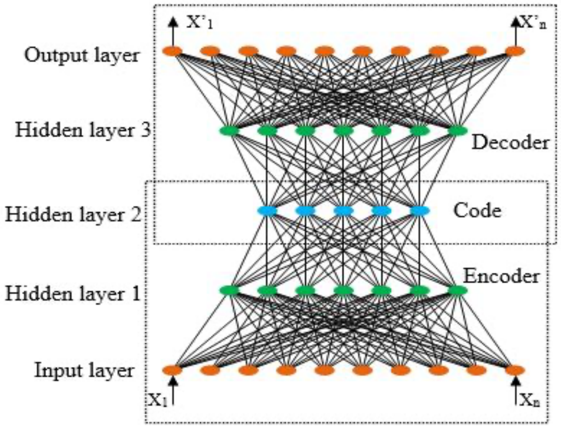

Fig. 3. Construction of Autoencoder

The autoencoder is made up of three parts: encoder, latent representation (code), and decoder. The encoder's role is to translate the input X (such as images, video, audio, and text) into latent space h. The decoder, on the other hand, use the input of latent space to reconstruct the output X'as similar to input data (i.e., $X^{(i)} = X'^{(i)}$). It uses n inputs and bias received at a single layer of the autoencoder's hidden layer. The latent representation in an autoencoder is represented by

$$c = f\left(\sum_{i=0}^{m} w_{ij} x_i + b_i\right) \qquad (1)$$

where w is the weight matrix of input values and hidden neurons, b is the bias of hidden neurons, and c is the perceptron of the network's hidden unit. Similarly, f(x) is a sigmoid function defined as follows:

$$f(x) = {}^{1}\!/_{1 + exp^{-z}} \qquad (2)$$

$$Z_i = \sum_{j=1}^{m} W_{ij} c + b_i \qquad (3)$$

$$x' = f(z_i) \qquad (4)$$

The output X' of the autoencoder is compared with desired outputs, x, and the mean squared error values have been calculated. Connected weights are adjusted according to weight changes. The autoencoder's main objective is to recreate its output as input. This is accomplished by using a customized loss function during autoencoder learning. This loss function is known as reconstruction loss, and it is often defined as the MSE (Mean Squared Error) between the input and output values. The loss function penalizes the network during training for producing outputs that vary from the input. The loss function is defined as follows:

$$Loss\ function = \|X - X'\|^2 \quad (5)$$

An autoencoder can be of different types based on construction, namely, vanilla autoencoder, undercomplete autoencoder, sparse autoencoder, denoise autoencoder, convolutional autoencoder, and variational autoencoder. A simple form of the autoencoder is vanilla which consists of a single-layer encoder and decoder. When the latent space neuron is lower than the input size, it is called an undercomplete autoencoder. Sparse autoencoder, stacked autoencoder, denoising autoencoder, and variational autoencoder are the regularized version of the vanilla autoencoder. Finally, the convolutional autoencoder is similar to the convolutional neural network used as unsupervised convolutional filters to reconstruct the images from higher dimensional to lower dimensional space.

III. AUTOENCODERS ON EMOTION RECOGNITION

Human-computer interaction (HCI) [36] has been a developing area of study in recent decades. Machines are often able to recognize human emotion through facial expressions. Table 1 shows the detailed study of autoencoder on facial emotion recognition. This study consists of factors related to year, variation of autoencoder, datasets, classifier, and accuracy of the proposed autoencoder. However, hardly a few studies on facial emotion recognition using autoencoders have been undertaken so far. A different regularized autoencoder can be used on the FER system and is divided into four categories such as deep autoencoder, sparse autoencoder, stacked autoencoder, and convolutional autoencoder. The first study was conducted in 2014 using the optical flow method in a sparse autoencoder [16]. From the below table, it is seen that residual autoencoder, stacked autoencoder, deep stacked autoencoder, sparse autoencoder, deep autoencoder, and convolutional autoencoder are the standardized variation of the vanilla autoencoder.

A. *Deep Autoencoder*

A deep autoencoder is constructed by two belief networks having two or three shallow layers. The first part is the encoder, and the other is the decoder part. It has more layers than a simple (Vanilla) autoencoder for extracting complex features. The increase of hidden neurons in the autoencoder can increase recognition accuracy. However, a large number of hidden layer nodes does not guarantee improved recognition accuracy. Sometimes it leads to a vanishing gradient problem. For this reason, sparsity has been imposed sometimes into the activation function. The deep autoencoder experiment [11][13] on the CK+ dataset shows the effectiveness of non-linear dimensionality reduction than linear PCA. It increased the learning capability of the network and achieved 99.60% accuracy.

B. *Sparse Autoencoder*

Sparse autoencoders enforce the sparsity constraints in the hidden layer. This can be added on before the latent space or instead of it. It learns some important information by using a large number of neurons in the latent space. There are two ways to apply sparsity constraint on autoencoder. The first one is L1 regularization which is directly applied instead of weights on activation. Another method is Kullback-Leibler (KL)-divergence method which measures the probability distribution between neurons. When the output is close to 1 or otherwise inactive, which means close to 0, it allows only a limited number of neurons to be triggered at the same time. This help to discover more interesting structure in the input data. The L1 regularization is most commonly used as a sparsity constraint. The experiment [29][12] showed the deep sparse autoencoder extracted more relevant information on the JAFFE dataset. However, the conventional facial feature extraction method in deep networks suffers from different emotional features in the training stage. The reason is that all the neuron is activated at the same time in a deep network. So, it could be difficult to distinguish the appropriate features for each emotion classification. In addition, the deep neural network often suffers from learning high-level features, local extrema, gradient diffusion problems, and computational complexity when increasing the data size.

Some authors [15][9][11] have used the non-linear machine learning method for extracting the features from the raw facial images. These experiments show the effectiveness of the holistic representation of features in the autoencoder for lower dimensionally. HOG, LBP, optical flow method, and fuzzy neural network have been used to extract the features when using the autoencoder for the FER system. This approach has helped the autoencoder to learn robust and discriminative features from feature vectors. This autoencoder has been trained and tested with CK+ and JAFFE datasets and achieved above 90% of accuracy.

C. *Stacked Autoencoder*

A stacked autoencoder is another variation of the layer-wise training method [2]. The output of one autoencoder is given to the input of another autoencoder. In simple words, stacking more than one autoencoder for reconstructing the output is similar to input with higher accuracy. Hinton and the PDP group [37] initially proposed the notion in the 1980s to solve the problem of backpropagation. It is similar to the deep belief network (DBN) which is composed of stacked DBN and its enhanced representation of learning. It is trained one by one and fine-tuned with backpropagation. To adjust the weights of the autoencoder, the gradient descent backpropagation method is utilised. It calculates the increased weight or decreased weight based on an increased weight by small values. This approach [2] has improved the facial emotion recognition accuracy to above 95% on JAFFE and CK+ datasets. However, the accuracy of the model has been varying based on the dataset still, because the facial expression is not universal. It is culturally specific.

D. Convolutional Autoencoder

A convolutional autoencoder is a kind of convolutional neural network [13]. The final part of the fully connected layer is replaced again by the de-convolution layer for reconstructing the output as input with a lower dimension, trying to reduce the reconstruction loss. The problem in the conventional network when using gradient descent, it suffers to find the local optimum and recall the training data due to more parameters. This complicates the quick generalization process of CNN network. After pre-training, the convolutional autoencoder has been fine-tuned by the final layer. Finally, the backpropagation method is used to fine-tune the entire network in a supervised manner. This method obtained over 90% of accuracy in RaFD, FER-2013, and JAFFE datasets.

E. Another Variation of Autoencoders

Many facial expressions are closely related to other emotions which leads to overlapping classes. For that reason, residual autoencoder [28] has been introduced recently for overlapping classes on feature space. This problem has been raised by an imbalanced class of images in the dataset. To address these issues affinity-based overlap reduction techniques have been introduced after the residual variational autoencoder trained with unlabeled datasets. The overlapped latent vector is then subjected to affinity-based approaches to decrease the overlapping across classes. In addition, it has been compared with many well-known classifier algorithms.

Generative adversarial autoencoder [30] is introduced to extract both local and global spatial information on facial images to reduce the parameter size of the network that uses facial symmetry. Furthermore, it has been trained with a greedy layer-wise learning algorithm. Finally, the fuzzy neural network-based autoencoder has been introduced for the identification of human facial expressions based on how the user felt inside. The fuzzy C-Mean approach was utilized to cluster the input data for this work, and a deep sparse autoencoder was constructed for emotion categorization. This approach achieved above 95% on KDFE and 81% on NAO facial expression datasets.

From the overall analysis, it is clear that the autoencoder shows superior performance on facial emotion images due to the feature reduction of higher dimensional data into the lower dimensional space. It also overcame the problems such as learning efficiency and computational complexity in conventional deep networks. In addition, the problem of gradient descent in backpropagation has been solved by a pre-trained autoencoder by layer-wise training method which increases the recognition rate and accuracy of existing network performance. Still, it needs a better improvement on FER for real-time emotion recognition.

TABLE 1. Related Work of Autoencoder on FER System

#	Author	Year	Variation of Autoencoder	Datasets	Classifier	Accuracy (%)
1	Chatterjee, Sankhadeep, et al. [28]	2022	Residual Autoencoder	Affectnet	LR, NB, RF, KNN, MLP, SVM, XGBOOST	LR: 92, NB: 93, RF: 92, KNN: 97 MLP: 87, SVM: 94, XGBOOST: 91
2	Lakshmi, Ponnusamy [15]	2021	Deep Stacked Autoencoder (HOG, LBP)	JAFFE, CK+	SVM	CK+: 97.66 %
3	Chen, Luefeng, et al. [29]	2021	Deep Sparse Autoencoder	JAFFE, CK+	Softmax regression	JAFFE: 90.12 CK+: 100.03
4	Allognon, Sevegni Odilon Clement, Alceu de S. Britto et al. [13]	2020	Deep Convolutional Autoencoder	FER-2013, RECOLA-2016	SVM	Valence: 0.516 Arousal: 0.264
5	Ruiz-Garcia, Ariel, et al.[30]	2020	Generative Adversarial Stacked Autoencoder, ResNet	KDFE, NAOfaces	CNN model with fine-tuned	KDFE: 98.07 NAOfaces : 81.36
6	Chen, Luefeng, et al.[31]	2020	Fuzzy neural network with Sparse Autoencoder	CK+, CASIA	NA	CK+: 81.31 CASIA:82.08
7	Chen, Luefeng, et al.[27]	2018	Softmax Regression based Deep Sparse Autoencoder	JAFFE, CK+	Softmax regression	JAFFE: 89.12 CK+: 98.03
8	Zeng, Nianyin, et al.[9]	2018	Deep Sparse Autoencoder (HOG & LPB)	CK+	NA	CK+: 95.79
9	Prieto, Luis Antonio Beltrán, and Zuzana Komínková Oplatková [10]	2018	Autoencoder with CNN	RaFD	NA	RaFD : 90
10	Usman, Muhammad, Siddique Latif, and Junaid Qadir [11]	2017	Deep Autoencoder (HOG)	CK+	SVM	CK+ : 99.60 PCA: 96.8
11	Dachapally, Prudhvi Raj [7]	2017	Autoencoder with Convolutional Neural Network	JAFFE, LFW	NA	JAFFE: 86.38 LFW: 67.62
12	Ruiz-Garcia, Ariel, et al. [8]	2017	Stacked deep Convolutional Autoencoder	KDFE	MLP	KDFE: 92.52
13	Majumder, Anima, Laxmidhar Behera et al. [28]	2016	Deep Neural Network-based Autoencoder	MMI, CK+	Kohonen Self-organized map (SOM)	MMI: 93.19 CK+ : 95.01
14	Huang, Binbin, and Zilu Ying [12]	2015	Sparse Autoencoder	JAFFE	NA	JAFFE: 94
15	Liu, Yunfan, et al. [16]	2014	Sparse Autoencoder (Optical Flow)	CK+,	Softmax	CK+: 91.6

IV. FACIAL EXPRESSION DATASETS

Datasets are crucial in the machine and deep learning techniques. Creating a better model requires a huge variety of datasets for proper training, validation, and testing. The facial emotion datasets have been divided into two categories, namely spontaneous datasets and posed datasets. Spontaneous datasets have been captured from natural environments whereas posed datasets have been captured from artificial facial expressions. Some commonly used facial emotion datasets on autoencoders are given below.

JAFFE [17]: The Japanese Female Facial Expression (JAFFE) dataset contains 213 photos of various emotions captured on the faces of 10 female individuals. These are seven expressions: happy, sad, surprise, neutral, fear, anger, and disgust. The resolution of the image size is 256 x 256 with associated labels. **CK+** [18] : Cohn-Kanade (CK) is the first 3D public dataset for facial emotion recognition. It contains 593 videos from 123 subjects aged ranging from 18 to 50 years. Each video has been recorded with 30 frames per second with resolutions of 640 x 490. Out of this, 327 are classified with seven expressions: happy, angry, fear, surprise, sad, and contempt. This dataset has been widely used in many facial emotion recognition models. Further, this database has been extended (CK+). **FER-2013** [20]: FER-2013 contains 35,685 RGB facial images with different expressions. The size of each image is restricted to 48 x 48 along with seven labels, namely, happy, surprise, fear, neutral, sad, angry, and disgust. **AffectNet** [21]: AffectNet is the largest facial emotion dataset collected from real-world images on the web. It consists of 0.4 million images with associated labels and also contains arousal and valence intensity of eight expressions such as happy, neutral, angry, sad, fear, contempt, disgust, and surprise. **KDFE** [22]: The Karolinska Directed Emotional Faces (KDEF) collection contains 4900 photographs of human facial expressions from 140 people aged 20 to 30. It consists of seven emotions such as happiness, fear, neutral, anger, sadness, surprise, and disgust. **MMI** [19]: The MMI collection includes 2900 videos and high-resolution photos from 75 different people. The resolution frame rate is 720 x 526. These are fully annotated with Action units (AUs) and partially coded with frame levels either neutral, apex, onset, or offset. It has both posed and spontaneous expressions. **RECOLA-2016** [23]: RE-mote COL-laborative and Affective (RECOLA) collection of audio-visual datasets divided into train, validation, and test sets. The frame rate of this dataset is 40ms. **NAOfaces** [24]: This dataset consists of 196 images collected in an unconstrained environment. It includes different ethnic groups of facial images. **RaFD** [25]: Radboud Face dataset (RaFD) consists of eight facial expressions of 67 subjects. It is freely accessible for non-commercial businesses and also for research purposes upon proper request. It contains 28,709 facial expression images. Each emotion was taken with three different looks and five camera angles. **LFW** [26]: The labeled faces in the wild (LFW) dataset contains 13,233 photos from 5,749 people and was developed to address the challenge of unconstrained face identification. It includes four different sets of LFW images and different types of aligned faces.

V. EMOTION CLASSIFICATION

A classification algorithm is a supervised learning technique used to categorize different classes based on training data. After the trained autoencoder, the encoder part is taken away and connected with classifiers like SVM [33], Softmax layer [35], and multi-layer perceptron [34]. The most commonly used classifiers with autoencoder are SVM and Softmax regression. All the extracted useful features from the raw data are stored in latent space which is connected with a classifier algorithm to categorize the trained observation and new upcoming observation. This kind of dimensionality reduction method helps the FER system to learn more relevant patterns of each emotion. In addition, the autoencoder and its variations are evaluated by mean squared error, accuracy, the influence of a number of hidden layers, and their recognition rate, confusion matrix, ROC curve, training, and testing time of the model. These metrics are used to validate the autoencoder for real-time emotion recognition in a complex environment.

VI. CONCLUSION

This paper examines a comprehensive review of facial emotion recognition on autoencoder. These kinds of autoencoders are introduced to address the issues of conventional feature selection methods and deep neural networks. In general, an autoencoder is a type of unsupervised learning technique that is extensively used to reduce high-dimensional input into lower-dimensional space. Finally, a non-linear classifier is used to train and test for the classification tasks. The main aim of this autoencoder is to learn high-level features while reshaping input data into low-level feature vectors based on network structure and ignoring some irrelevant features from input data. Compared to the conventional feature extraction method, the autoencoder efficiently extracts the most useful information in an unsupervised manner. So far, merely a few experiments on facial emotion recognition using an autoencoder have been conducted. Apart from facial emotion recognition, autoencoder has been used in many applications and achieved remarkable results both in supervised and unsupervised learning problems.

Many researchers have introduced variations of different autoencoders to solve specific problems such as the complexity of the network, learning efficiency, overlapping issues, local extrema, and gradient diffusion problems in conventional methods. This paper reviewed all the available works using autoencoder on the FER system between 2014 to 2022. CK+ and JAFFE datasets are widely tested datasets on autoencoder and achieved over 90% accuracy and have been deployed in human-robot interaction.

ACKNOWLEDGMENT

This work has been supported by the Centre for Machine Learning and Intelligence (CMLI) sponsored by the Department of Science and Technology (DST-CURIE).

REFERENCE

[1] J. Wang, H. He, and D.V. Prokhorov, "A folded neural network autoencoder for dimensionality reduction". *Procedia Computer Science*, *13*, 2012, 120-127.

[2] Y Wang, H Yao, S Zhao, and Y Zheng, "Dimensionality reduction strategy based on auto-encoder", In *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service,*2015**,** pp. 1-4.

[3] P. Viola, and M. J. Jones, "Robust real-time face detection", *International journal of computer vision*, *57*(2), 2004,137-154.

[4] M.F. Valstar , B. Jiangm, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge", In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG),* 2011, pp. 921-926. IEEE.

[5] A. Daffertshofer, C.J. Lamoth, O.G. Meijer, and P.J.Beek, "PCA in studying coordination and variability: a tutorial", *Clinical biomechanics*, *19*(4), 2004, 415-428.

[6] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with the learning-based descriptor". In *2010 IEEE Computer society conference*

[7] P. R. Dachapally, "Facial emotion detection using convolutional neural networks and representational autoencoder units", *arXiv preprint arXiv:1706.01509*. 2017

[8] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade, "Stacked deep convolutional auto-encoders for emotion recognition from facial expressions", In *2017 International Joint Conference on Neural Networks (IJCNN),2017,* pp. 1586-1593. IEEE.

[9] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A.M.. Dobaie, "Facial expression recognition via learning deep sparse autoencoders", *Neurocomputing*, *273*, 2018, 643-649.

[10] L.A.B. Prieto, and Z.K. Oplatková, "Emotion recognition using autoencoders and convolutional neural networks", In *Mendel* (Vol. 24, No. 1, 2018, pp. 113-120.

[11] M. Usman, S. Latif, and J. Qadir, "Using deep autoencoders for facial expression recognition", In *2017 13th International Conference on Emerging Technologies (ICET),2017,* pp. 1-6. IEEE.

[12] B. Huang, and Z. Ying, "Sparse autoencoder for facial expression recognition", In *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom),2015,* pp. 1529-1532. IEEE.

[13] S.O.C. Allognon, A.D.S. Britto, and A.L. Koerich, "Continuous Emotion Recognition via Deep Convolutional Autoencoder and Support Vector Regressor", In *2020 International Joint Conference on Neural Networks (IJCNN),2020,* pp. 1-8. IEEE.

[14] Y. Liu, X. Hou, J. Chen, C. Yang, G. Su, and W. Dou," Facial expression recognition and generation using sparse autoencoder", In *2014 International Conference on Smart Computing*, 2014, pp. 125-130. IEEE.

[15] D. Lakshmi, and R. Ponnusamy, "Facial emotion recognition using modified HOG and LBP features with deep stacked autoencoders", *Microprocessors and Microsystems*, *82*, 2021, 103834.

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

on computer vision and pattern recognition, 2010, pp. 2707-2714. IEEE**.**

[17] The Japanese Female Facial Expression Database. Available Online: https://www.kasrl.org/jaffe_download.html (accessed on 29 Sep 2022)

[18] The Extended Cohn-Kanada Database. Available online: https://www.ri.cmu.edu/ (accessed on 29 Sep 2022)

[19] MMI Database. Available online: https://mmifacedb.eu/

[20] Facial Expression Recognition 2013 Database. Available Online: https://www.kaggle.com/datasets/msambare/fer2013

[21] AffectNet Database. Available online: http://mohammadmahoor.com/affectnet/

[22] Karolinska Directed Emotional Faces Database. Available Online: https://www.kdef.se/

[23] Recola datasets. Available Online: https://diuf.unifr.ch/main/diva/recola/download.html

[24] A. Ruiz-Garcia, N. Webb, V. Palade, M. Eastwood, and M. Elshaw, "Deep learning for real-time facial expression recognition in social robots", In *International Conference on Neural Information Processing,* pp. 392-402. Springer, Cham.

[25] Radboud Faces Database. Available Online: https://www.ru.nl/bsi/research/facilities/radboud-faces-database

[26] Labeled Faces in the Wild. Available Online: http://vis-www.cs.umass.edu/lfw/

[27] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota,"Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction", *Information Sciences*, *428*, 2018, 49-61.

[28] S. Chatterjee, A.K. Das, J. Nayak, and D. Pelusi, "Improving Facial Emotion Recognition Using Residual Autoencoder Coupled Affinity-Based Overlapping Reduction", *Mathematics*, *10*(3),2022, 406.

[29] L. Chen, M. Wu, W. Pedrycz, and K. Hirota, "Deep sparse autoencoder network for facial emotion recognition", In *Emotion Recognition and Understanding for Emotional Human-Robot Interaction Systems,* 2021, pp. 25-39. Springer, Cham.

[30] A. Ruiz-Garcia, V. Palade, M. Elshaw, and M. Awad, "Generative adversarial stacked autoencoders for facial pose normalization and emotion recognition", In *2020 International Joint Conference on Neural Networks (IJCNN), 2020,* pp. 1-8. IEEE.

[31] L. Chen, W. Su, M. Wu, W. Pedrycz, and K. Hirota, "A fuzzy deep neural network with sparse autoencoder for emotional intention understanding in human-robot interaction", *IEEE Transactions on Fuzzy systems*, *28*(7), 2020, 1252-1264.

[32] M. Gogoi, and S.A. Begum, "Image classification using deep autoencoders", In *2017 IEEE international conference on computational intelligence and computing research (ICCIC)*, 2017, pp. 1-5. IEEE**.**

[33] V. Jakkula," Tutorial on support vector machine (svm)", School *of EECS, Washington State University*, *37*(2.5), 2016, 3.

[34] Taud, H., and Mas, J. F, "Multilayer perceptron (MLP)", In *Geomatic approaches for modeling land change scenarios*, 2018, pp. 451-455. Springer, Cham.

[35] A. Joulin, M. Cissé, D. Grangier, and H. Jégou, "Efficient softmax approximation for gpus", In *International conference on machine learning*, 2017, pp. 1302-1310, PMLR.

[36] Harper, E. R., Rodden, T., Rogers, Y., Sellen, A., and Human, B. "Human-Computer Interaction in the year 2020",2018.

[37] Baldi, P. "Autoencoders, unsupervised learning, and deep architectures". In *Proceedings of ICML workshop on unsupervised and transfer learning*, 2012, pp. 37-49. JMLR Workshop and Conference Proceedings.

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