

Deep learning approaches for recognizing facial emotions on autistic patients

Fatima Ezzahrae El Rhatassi¹, Btihal El Ghali¹, Najima Daoudi^{1,2}

¹ITQAN Team, LYRICA Lab, Information Sciences School (ESI), Rabat, Morocco

²SSLab, ENSIAS, Mohammed V University, Rabat, Morocco

Article Info

Article history:

Received Jan 19, 2024

Revised Apr 4, 2024

Accepted Apr 16, 2024

Keywords:

Autism

Chatbot

Convolutional neural network

Facial emotion recognition

Vision transformers

ABSTRACT

Autistic people need continuous assistance in order to improve their quality of life, and chatbots are one of the technologies that can provide this today. Chatbots can help with this task by providing assistance while accompanying the autistic. The chatbot we plan to develop gives to autistic people an immediate personalized recommendation by determining the autistic's state, intervening with him and build a profile of the individual that will assist medical professionals in getting to know their patients better so they can provide an individualized care. We attempted to identify the emotion from the image's face in order to gain an understanding of emotions. Deep learning methods like convolutional neural networks and vision transformers could be compared using the FER2013. After optimization, conventional neural network (CNN) achieved 74% accuracy, whereas the vision transformer (ViT) achieved 69%. Given that there is not a massive dataset of autistic individuals accessible, we combined a dataset of photos of autistic people from two distinct sources and used the CNN model to identify the relevant emotion. Our accuracy rate for identifying emotions on the face is 65%. The model still has some identification limitations, such as misinterpreting some emotions, particularly "neutral," "surprised," and "angry," because these emotions and facial traits are poorly expressed by autistic people, and because the model is trained with imbalanced emotion categories.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Fatima Ezzahrae El Rhatassi

ITQAN Team, LYRICA Lab, Information Sciences School (ESI)

Rabat, Morocco

Email: fatima-ezzahrae.el-rhatassi@esi.ac.ma

1. INTRODUCTION

Everyday life includes having conversations on a daily basis. Since societies are made up of individuals, communication is crucial. Speaking, gesturing, making hand motions, and expressing emotions, can all help. Making your message understood by the other person is the most important. There are people who are nonverbal, have illnesses like autism, which find it difficult to communicate with others, or would rather live alone than in a community where they risk being misinterpreted or criticized. Autism-related patients need extra guidance, support, and care from family, caregivers, and educators in order to feel better, take care of their health, and see more than just their home as a place where they are accepted and safe from the scrutiny of others [1], [2].

Autism as a developmental abnormality, makes it difficult for autistic people to interact, socialize, and live with others. In addition to medical treatment, which may include behavioral, cognitive, and even genetic testing, emotional support is necessary in order to understand their mental state and offer help [3].

For this reason, we considered utilizing a chatbot that can detect and recognize the user's emotions in order to respond appropriately, inform parents of their child's mental state, and avert situations or problems that may arise from his negative mood and behavior.

We see that in order to create a chatbot, it is essential to first capture the user's facial emotion and comprehend his mental state. This way, the chatbot can build conversations with the user based on his mood and attempt to help him feel better when he's not feeling well until the patient can receive physical assistance. For example, if the chatbot detects signs of distress in the user's facial expressions, it can provide supportive messages or suggest relaxation techniques to alleviate his discomfort. Additionally, understanding the user's mental state can also enable the chatbot to tailor its responses and recommendations to suit his individual needs and preferences.

We will subsequently employ an array of artificial intelligence methods, including vision transformer (ViT) and convolutional neural network, to assess the efficacy of these approaches in discerning the autistic individual's present mental state through their facial imagery. By leveraging advanced artificial intelligence (AI) technologies such as the vision transformer and convolutional neural network, we aim to accurately analyze subtle facial cues and expressions that may indicate the individual's emotional state. Moreover, through rigorous testing and evaluation, we seek to determine the reliability and effectiveness of these methods in providing insights into the emotional well-being of autistic individuals, paving the way for more personalized and responsive support systems.

Deep learning techniques are increasingly being used after the coronavirus disease 2019 (COVID-19) pandemic in image processing and categorization, in particular the recognition of human emotions and facial features to understand and improve the psychological state of people [4]–[6], or simply to understand public perception of a given subject in social networks, since the latter are much more widely used than before [7]–[9]. Various approaches exist for identifying facial emotions, such as employing machine learning or deep learning techniques like transformers and neural networks. Specifically, conventional neural networks (CNNs), a kind of deep learning method, have demonstrated potential in recognizing facial emotions. Notably, CNNs have proven effective in simplifying tasks during the feature extraction and preprocessing phases of this process.

Talaat [10] employed CNN for the identification of emotions from images featuring the faces of children with autism. Her approach involved optimizing hyperparameters, followed by training with the Inception-ResNetV2 model. Subsequently, she utilized the U-net model for segmentation to enhance the speed and precision of detection. The model achieved an impressive 99% accuracy, although it was limited by a small dataset: 1,200 images for training and 220 for testing. Plans are in place to expand the dataset in future studies.

As a result of the lack of research conducted in this area specifically for autistics, we chose to expand our analysis to include additional datasets that already exist and have been the focus of numerous studies and technical developments related to facial emotion identification. From these datasets, we find facial expression recognition 2013 (FER-2013), The extended Cohn-Kanade dataset (CK+) [11], AffectNet [12], real-world affective faces database (RAF-DB), and Japanese female facial expression (JAFFE) [13]. FER-2013 is a sizable collection of 35,587 grayscale portraits. It can therefore be utilized for training deep learning models. Additionally, the photos in FER-2013 are taken from the internet and are in close proximity to real-world settings, in contrast to datasets that comprise posed expressions by actors (such as CK+ and JAFFE). Additionally, the dataset was utilized in a competition. It has been thoroughly examined and benchmarked in a number of research projects, enabling efficiency analyses and technique comparisons. Finally, with each image on FER-2013 featuring a single person's face that has been cropped and centered [14], [15], providing a straightforward dataset structure for emotion classification tasks, which is precisely what our study aims to provide: the ability to recognize emotion from a face's photo; from an autistic's photo.

The dataset FER2013 is made up with 35,587 48×48 greyscale photos of faces classified into seven categories-surprise, anger, disgust, fear, happiness, sadness, and neutral. It can be accessed online. Furthermore, the seven groups are not equal; only 436 images fall into the category of disgust, while 7,215 images fall into the category of happiness [16]. This imbalance in the dataset may pose challenges for training and evaluating machine learning models, as it can affect the model's ability to accurately recognize and classify some emotions.

Goodfellow *et al.* [16] achieved an accuracy of 65.5% through the use of convolutional neural networks and hyper-parameter optimization. Zhang *et al.* [17] reported that they were able to optimize the model generated by adding more features, such as facial landmarks, and using external datasets to augment the data, resulting in a 75.1% accuracy rate. In order to achieve accuracy between 72.81% and 73.53%, some of the works combined multiple models, such as local (multi) head channel (LHC), with a pre-trained backbone, the ResNet34v2 by Pecoraro [18].

In 2019, Georgescu *et al.* [19] developed a novel model, attaining an accuracy of 75.42% through the incorporation of additional CNNs, the bag of visual words (BOVW) model, linear support vector

machine (SVM), and various local learning methods. CNNs that have been pretrained as ResNet or VGG have been used in a lot of recent work; these models can identify larger features in images because they have been trained in datasets that contain, for instance, over 3 million images.

When it comes to transformers, they first emerged in natural language processing [20], where they revolutionized the field by introducing the encoder-decoder architecture. This architecture allows the transformer model to split the input phrase into sequences of words, analyze and understand each part simultaneously, and capture important contextual information while processing. By leveraging self-attention mechanisms, transformers excel at capturing long-range dependencies and contextual nuances, enabling them to generate coherent and contextually relevant responses. Recently, there has been a growing interest in extending the use of transformers beyond text analysis to other domains, including image processing. Specifically, researchers have started exploring the application of transformers in image analysis tasks, such as facial emotion recognition.

Dosovitskiy *et al.* [21] and associates used transformers for image recognition. An image classification method that directly leverages the transformer architecture was proposed. Multiple patches are created from the input image by the vision transformer (ViT) model. After that, each patch is flattened and projected linearly into a single, constant dimension. The position of each picture patch in the image is then displayed by adding a position embedding to it. They also employ the conventional method of classifying data, which involves appending an additional learnable "classification token" to the sequence. The recommended model performs better than 90%, however it was tested on datasets that increase in size: JFT-300M, ImageNet, and ImageNet-21k. In the study involving transformers on the FER2013 dataset, it was merged with other datasets like AffectNet and CK+48 for the purpose of data augmentation, increasing the total size of the final datasets [22]. The data was then categorized into test, validation, and training sets for evaluating their AVFER model. The AVFER model (incorporating AffectNet, FER-2013, and CK+48) was applied to various models including ViT-B/16/S, ViT-B/16/SAM, ViT-B/16/S, and ResNet-18. The results showed accuracy rates of 50.05%, 52.25%, 52.42%, and 53.10%, along with AUC values of 0.843, 0.837, 0.801, and 0.589, respectively. Based on the survey in this section, Table 1 showcases a selection of cutting-edge results applied to the FER2013 dataset.

This table seems to indicate that the two artificial intelligence techniques that are expressed on CNN and transformers are more motivating when it is applied to achieve the facial emotion recognition, particularly when data augmentation is used and other pretrained models on larger datasets are combined. CNNs and transformers have demonstrated their ability to extract meaningful features from facial images and capture complex patterns, leading to more accurate emotion recognition results. Building upon these findings, we plan to leverage the capabilities of CNNs and vision transformers in our research on facial emotion recognition in individuals with autism.

To ensure coherence, the organization of this paper is outlined as follows: section 2 will present the suggested work. Section 3 offers an application of the CNN method to the autism dataset and the result will be examined and discussed. The concluding section will cover the effectiveness of this research, offering a roadmap for ongoing and future studies.

Table 1. Facial emotion recognition works applied to the FER2013 dataset

Variable	Accuracy	Method	Model improvement
CNN [17]	65.5%	CNNs	- Hyper-parameter optimization
	75.1%		- Adding more features (facial landmarks)
CNN [18]	72.81%	CNN + Pretrained neural networks	- Using external datasets to augment the data
	73.53%		- Multiple models are combined. LHC, with a pre-trained backbone, the ResNet34v2
CNN [19]	75.42%	CNN + Pretrained neural networks ResNet or VGG	- Using additional CNNs, the BOVW model, linear SVM, and some local learning techniques
Vision transformer [22]	50.05%	ViT transformer + other pretrained models	- Data augmentation: The FER2013 dataset was combined with other datasets, (AffectNet and CK+48)
	52.25%		- AVFER (AffectNet, FER-2013, and CK+48) on the ViT-B/16/S, ViT-B/16/SAM, ViT-B/16/S, and ResNet-18 models.
	52.42%		
	53.10%		

2. METHOD AND APPROACH

Our work involves enabling personalized medicine for autistic patients through the development of a chatbot that will function as an auxiliary assistance to the autistic until a family member, caregiver, or other close relative arrives. The goal of this work is to get the autistic user talking and expressing himself freely

with a chatbot that attempts to identify his mood or state of mind, attempts to intervene if he becomes angry or encounter a situation that calls for assistance, and alerts his parents or caregiver. The user can message and communicate with the chatbot by text, video, or audio only [23].

To begin with, the decision to prioritize emotion recognition stems from the crucial role emotions play in communication and social interaction, particularly for individuals on the autism spectrum. As depicted in Figure 1, our approach involves leveraging facial images containing the faces of individuals with autism to detect and analyze emotions. By accurately recognizing emotions, we aim to gain insights into the emotional states of individuals with autism, enabling us to better understand their mood and facilitate meaningful interactions. Our methodology involves using facial images as input to our system, which will employ machine learning algorithms to recognize emotions and identify the user's mood. Additionally, we intend to consolidate information about emotions and situational contexts to generate appropriate interventions and conversations tailored to the individual's needs and preferences.

Technological aids as chatbots are among the most important innovations that have improved the lives of individuals with autism. Beyond the support these chatbots offer in providing assistance to the autistic person; this is the case that will be realized in our work. It can also assist in the autistic practice, increase his social skills, and remove him from the social exclusion and unfavorable social judgment that he may experience outside of his home [24].

To achieve emotion detection from facial images, it is essential to leverage advanced deep learning techniques, as highlighted by the related research. Both convolutional neural networks (CNNs) and Transformers have emerged as prominent tools in this domain, each offering unique strengths and capabilities for facial emotion recognition. Table 2 provides a comprehensive overview of the strengths of these artificial intelligence techniques in image analysis, underscoring their utility in facial emotion recognition tasks. By considering the insights gleaned in this table, we aim to leverage the complementary strengths of CNNs and transformers in our research to develop a robust and accurate emotion detection system tailored to the nuances of facial expressions in individuals with autism.

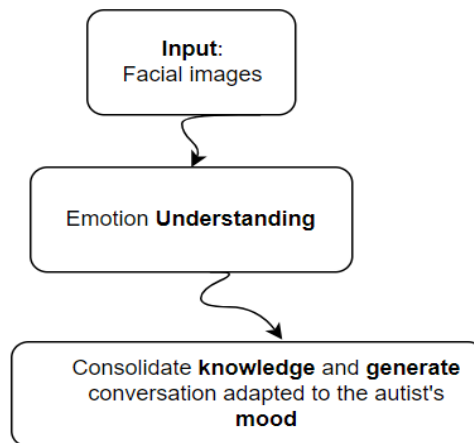


Figure 1. Focus on the chatbot's conception based on the identification of the autistic facial emotion

Table 2. Comparative table between deep learning techniques (CNN and vision transformer)

Technique	CNNs	Vision transformer
Characteristics of the technique's application	Specialized in image processing; using input images to learn the spatial hierarchies of features [25] Specialized in local feature identification as edges, textures and shapes [26] Robustness on variations in input as lightning changes, orientation or scale [28] Solid research area: CNNs applications on image recognition (facial) are numerous which provides a solid foundation for further development [30]	Effective capturing global dependencies in the data and good understanding on how facial features extract to convey an emotion [21] Utilization of an attention mechanism to concentrate on the facial areas most pertinent to emotion identification [27] Handling sequential data: well-suited for understanding the temporal dynamics of facial expressions (video's example) [29] Innovative Research Area: it is a relatively new area of research, and it offers the potential for novel approaches and breakthroughs in emotion recognition [22]

CNN is employed for its proficiency in feature extraction, a task carried out by its convolutional layers. Following this, the network's classification layers take on the role of recognizing emotions. Thus, to realize emotion recognition, the convolutional neural network is the first option to target. CNN are deep learning algorithms that, given an input-a face photo in this case-can learn to prioritize various features, objects, and details in the image and distinguish between them.

As illustrated in Figure 2, a typical CNN can be generally as a series of layers arranged sequentially: convolutional layers come first, succeeded by an activation function, then a pooling layer, and finally a fully connected layer. This sequence is repeated several times. The sequential construction of these layers enables CNN to acquire hierarchical characteristics by piling numerous hidden layers on top of one another in a predetermined order [25].

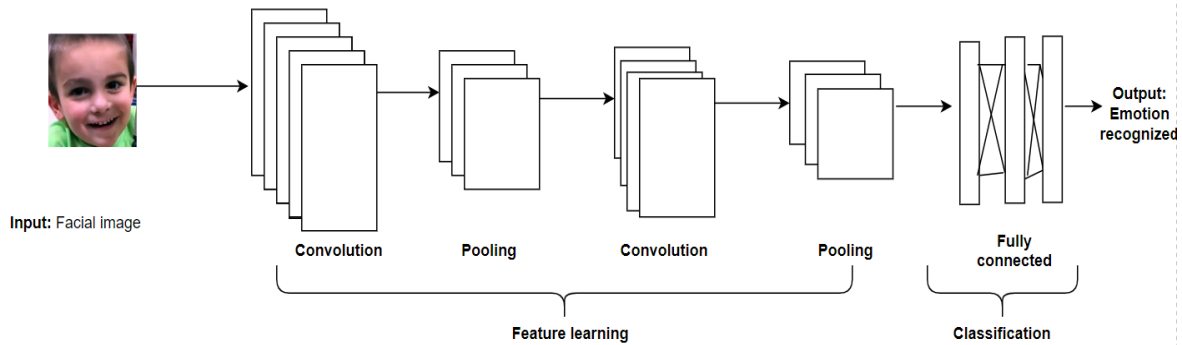


Figure 2. The architecture of a basic convolutional network

Typically, in a neural network, the fully connected layers are responsible for linking the extracted features to the end result, such as identifying facial emotions. On the other hand, the convolutional and pooling layers mainly handle the extraction of these features. As a result of repeatedly combining these procedures, the first layer learns to identify basic features in a picture, such as edges, while the second layer starts to identify more complex features [31]. By the last layer, the convolutional neural network can identify more complicated forms, such as classifying facial features and emotions.

As convolutional neural network techniques have demonstrated promising results in previous research, achieving accuracy rates of up to 75% with the incorporation of extra datasets and pretrained models [19], we have opted to incorporate them into the suggested design outlined in Figure 3. CNNs are well-suited for tasks involving image analysis, including facial emotion recognition, due to their ability to effectively capture spatial features and hierarchical representations. However, considering there is not currently a lab-tested dataset for autistic faces, we will be using FER2013 for the time being as a suitable starting point for our research.

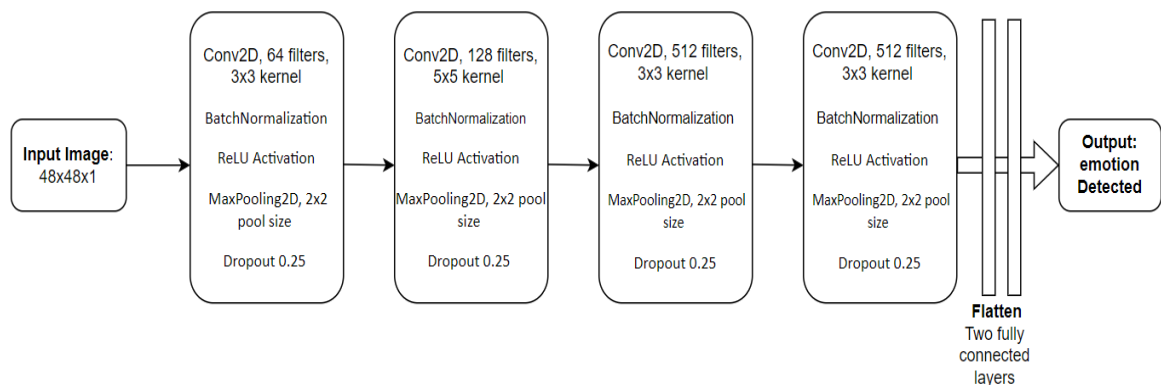


Figure 3. The proposed CNN model's architecture

2.1. CNN to FER-2013 dataset

This study focused on the analysis of seven distinct emotions: angry, disgust, fear, happy, sad, surprise, and neutral facial expressions. By examining these emotions, the study aimed to gain insights into the nuanced range of emotional states conveyed through facial expressions. The research methodology encompassed two primary stages: first, the extraction of facial features from the images, and second, the categorization and identification of these features into specific emotional states. This approach enables us to systematically analyze and interpret the facial expressions captured in the dataset, contributing to a deeper understanding of human emotion perception and expression.

2.2. Data augmentation

We started by augmenting the training dataset to include modified versions of the images. This approach aims to bolster the generalization abilities of the CNN model by exposing it to a broader spectrum of training examples. By using ImageDataGenerator in Keras, which is a tool that allows for real-time data augmentation as part of the image preprocessing pipeline. It helps in increasing the diversity of the training set by applying random but realistic transformations, such as rotation range, width shift range, height shift range, horizontal flip, and vertical flip.

2.3. Architecture of CNN model

The architecture of the model consists of four consecutive convolutional blocks. Each block includes a Conv2D layer, succeeded by batch normalization, rectified linear unit (ReLU) activation, max pooling, and finally, a dropout layer. After flattening the output of these convolutional blocks, the model incorporates two dense (fully connected) layers, each with batch normalization, ReLU activation, and dropout. The final part of the model is designed for multi-class classification and features a dense layer with seven units, utilizing softmax activation. This schema provides our model's architecture.

When utilizing this CNN model, it is learning and improving its performance on the training data, as evidenced by decreasing loss and increasing accuracy and precision. However, the increase in validation loss and decrease in validation accuracy and precision from epoch 298 to 299 might indicate that the model is beginning to overfit. This means it is getting better at memorizing the training data, but its performance on unseen data is getting a little unclear.

2.4. Metrics

2.4.1. Accuracy

Figure 4 shows the variation of accuracy and loss over epochs. The sharp rise in accuracy during the initial epochs of the model on the training datasets suggests rapid learning. After that, the model began to converge and make small gains and reaches 73%. Regarding the validation dataset, which was not viewed by the model during training. The line increases quickly and initially follows a similar trend to the training accuracy. Nonetheless, the validation accuracy reaches a peak sooner and stays quite constant after that. In the beginning, there is not much of a difference between the training and validation accuracy lines, typically indicating effective generalization by the model. The gap does, however, slightly widen, which may suggest that overfitting is possible. In terms of loss, the model predicts a drop in loss with time, signifying growth and learning and approaches to 0.96. The validation loss exhibits small variations towards later epochs, indicating a potential overfitting of the model.

2.4.2. Precision

With a precision of 0.8081%, the model demonstrates a high level of accuracy in predicting the positive class, achieving approximately 80.81% precision. This indicates that the model correctly identifies positive samples a significant portion of the time, reflecting its effectiveness in distinguishing between positive and negative instances. Furthermore, the observed improvement in precision suggests that the model is continually refining its predictive capabilities, particularly in avoiding mislabeling negative samples as positive. This trend underscores the model's adaptive nature and its potential for enhanced performance over time as it learns from additional data and iterations.

2.4.3. Confusion matrix

In Figure 5, the confusion matrix provides insight into the performance of the CNN model in classifying emotions using a dataset with known real values. The matrix reveals that the model exhibits strong performance for certain emotions, such as 'happy' and 'neutral', as indicated by high accuracy rates and low misclassification errors. However, the model appears to struggle with accurately classifying other emotions, such as 'disgust' and 'fear', as evidenced by higher rates of misclassification and lower accuracy scores.

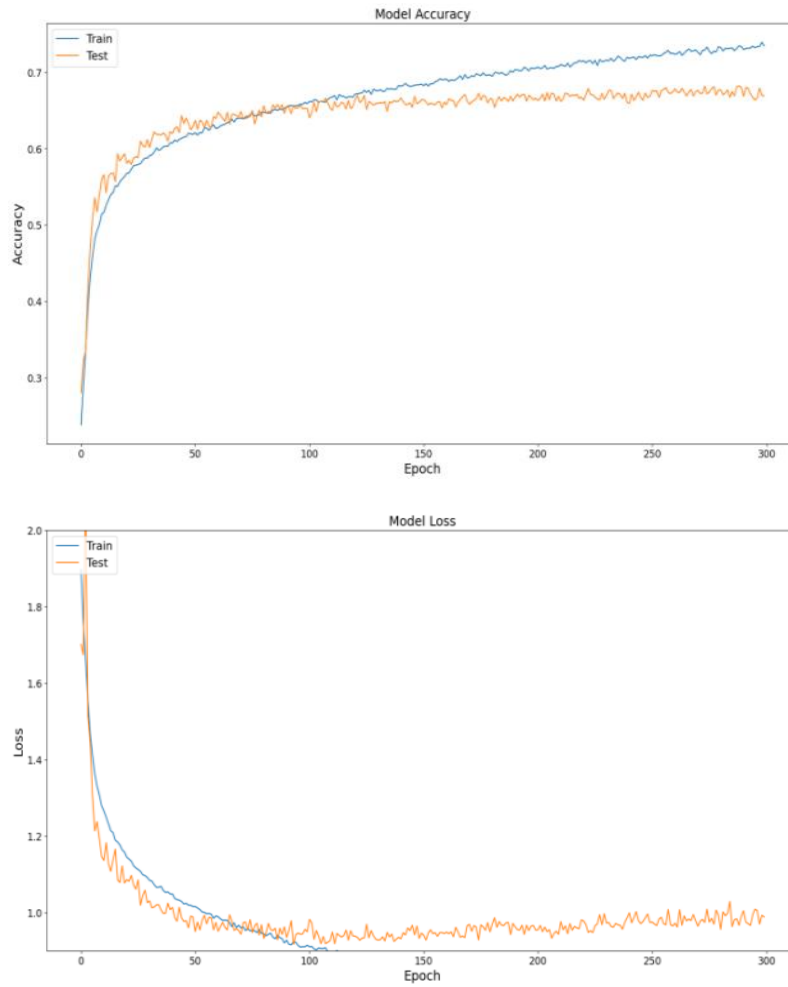


Figure 4. The variation of accuracy and loss over epochs during the training and validation (test) phases of the CNN

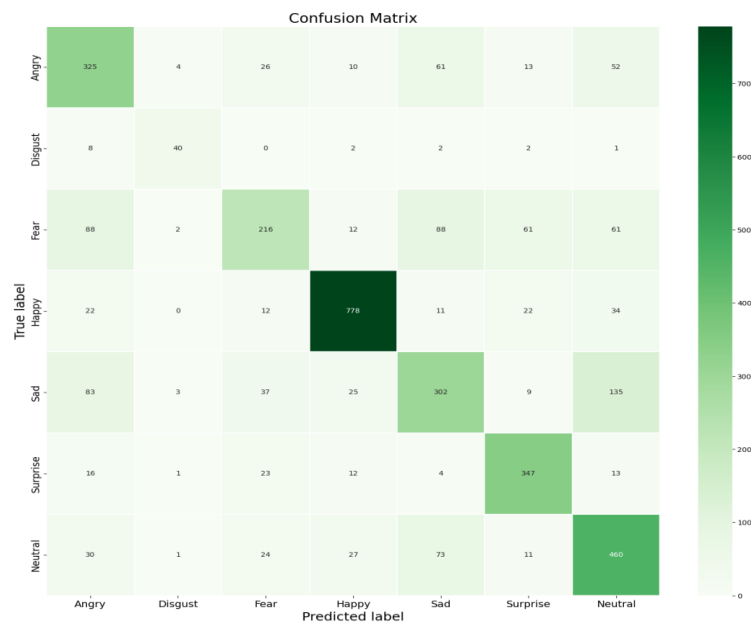


Figure 5. CNN confusion matrix of predicted emotions

2.5. Optimizing the CNN hyperparameters

To build a better model of our conventional neural network model, we must optimize the hyperparameters and select the optimal ones. We attempted to utilize the Bayesian method, a probabilistic model that examines the relationships between the hyperparameters. Since it would be laborious to examine every possible combination while training the model on a dataset, we considered selecting the following hyperparameters as our target values: Studying all the possible values for these hyperparameters will not require as much time and effort as it would if *learning_rate* is within $1e-4$ to $1e-2$, *dropout_rate* from 0.1 to 0.8, *num_filters* at either 16, 32, or 64, and *dense_size* set at 64, 128, or 256. After the optimization of the hyperparameters is done, the best values are *learning_rate*= $3.53956328e-03$, *dropout_rate*= $1.88596157e-01$, *num_filters*= $3.20000000e+01$ and *dense_size*= $1.28000000e+02$.

Over these epochs, the training loss slightly fluctuates but generally remains under 0.71. This indicates the model's predictions are getting more accurate with respect to the training data. The training accuracy also fluctuates but stays within the range of about 73.92% to 74.88%. As training continues, the model's performance on the validation set is not enhancing, as indicated by the instability and minor increase in validation loss, along with the oscillations in validation accuracy, suggesting overfitting. The model is not generalizing well to unseen data, which is a common challenge in machine learning.

2.6. Vision transformer

The principal breakthrough of the vision transformer lies in its use of self-attention mechanisms on image patches, allowing the model to evaluate the relative importance of different segments of the image. Essentially, this approach allows the model to recognize and understand the global interdependencies between these patches, allowing it to learn contextual relationships within the image. It concludes with a classification head that predicts the image's class.

To begin this step, we preprocessed the input data to match the format that the model expects, normalizing the pixel values and, if needed, rearranging the color channels as it is shown in Figure 6. The ViT model is a pretrained deep learning model on ImageNet-21k, a dataset that includes 14 million images spanning 21,000 classes. Thus, we installed the hugging face 'transformer' library next, which is made to load and process datasets quickly and easily. We also imported the ViTFeatureExtractor class and other visualization libraries and the transformers library.



Figure 6. Preprocessed images from FER2013 dataset

The preprocessing step consists of preparing the input into a vision transformer model. The input data is being preprocessed using ViTFeatureExtractor, which resizes every image to the resolution that the model expects; 224×224 and normalizes the format of the images into (channels, height, width). Then, the dataset is ready to be used as all the images have the same size and contrast as the Figure 7 illustrates. The FER2013 data is then fed into the model, and finally, the linear classifier-a layer neural network-maps the high-dimensional [CLS] token representation to the number of target classes. In this case, the number of classes equates to the seven emotions.

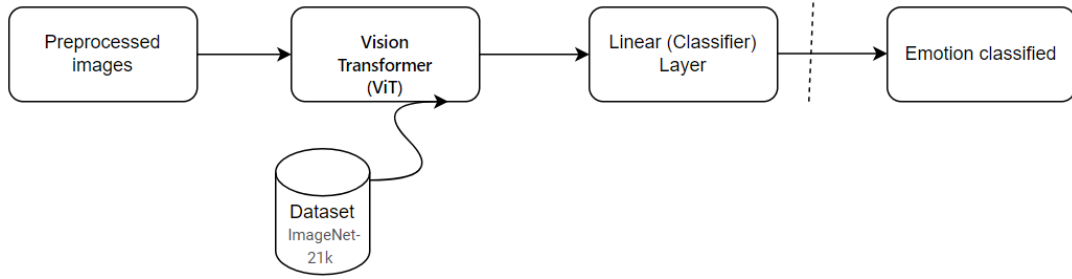


Figure 7. The application of ViT on FER dataset

2.7. Metrics

2.7.1. Accuracy

The model's classification accuracy on the validation dataset reflects the percentage of instances that have been correctly categorized. The accuracy improves over the epochs, reaching its peak at epoch 5 with 69.35%. However, it slightly decreases, which, combined with the increasing validation loss, might indicate overfitting.

2.7.2. Confusion matrix

The diagonal elements in Figure 8 are, as usual, the samples that were accurately predicted. Here, out of the 3,589 samples, 2,518 samples were accurately predicted. Therefore, 70% is the overall accuracy. The matrix indicates that the model is quite good at identifying happiness, which has the highest number of correct predictions and relatively few misclassifications. Surprise and neutral also have a high number of correct predictions. However, disgust is the most challenging emotion for the model to identify correctly, followed by anger and fear, which are often confused with each other. The model also struggles sometimes with distinguishing between fear and sadness. While the CNN and ViT models perform well in predicting happiness, surprise, and neutrality, they struggle to identify certain emotions like anger and fear. This could be caused by a number of things, such as the dataset's imbalanced number of photos for each emotion and an inherent difficulty in differentiating between specific emotions.



Figure 8. Confusion matrix of ViT to FER2013 dataset

3. THE CNN MODEL'S RESULTS APPLIED TO THE AUTISM DATASET AND DISCUSSION

In this work, we attempted to integrate various data sources to obtain images of autistic people in order to identify their moods. The dataset we utilized comprised 1,463 photos of faces of individuals with

autism, which were obtained from the Kaggle website [32], which included both autistic and non-autistic images. Their data was collected with the intention of detecting autistic from non-autistic photos. Since we were limited to detect the emotion of the autistic children in our instance, we used the CNN model in conjunction with the 719 images of autistic children found at Kaggle too [33]. Thus, this study included 2,192 images of autistic people's faces in total. Following the application of the CNN model described in the preceding section, the images are divided into these groups of emotions as it is shown in Table 3 (happy: 1,235, neutral: 796, angry: 71, sad: 39, surprise: 30, fear: 14, disgust:7).

Table 3. Distribution's pictures in dataset based on detected emotions

Emotions	Number of pictures
Angry	71
Disgust	7
Fear	14
Happy	1,235
Neutral	796
Sad	39
Surprise	30
Total	2,192

Facial emotion recognition will help us in this project to track the emotions of autistics, intervene, when necessary, warn others around them, and comfort them if they are upset or depressed. While applying to the dataset the CNN model, it appears from examining every image that the model performs admirably, achieving more than 65% accuracy. The confusion that really exists, though, is that the features of the autistic face can occasionally be misinterpreted in the real life not only by applying the model, particularly when it comes to the contrast between neutral and surprised or neutral and angry as it appears in Figure 9. We tend to assume that the autistic face is angry because the hands are covering the ears, even when the other facial qualities are normal and neutral.

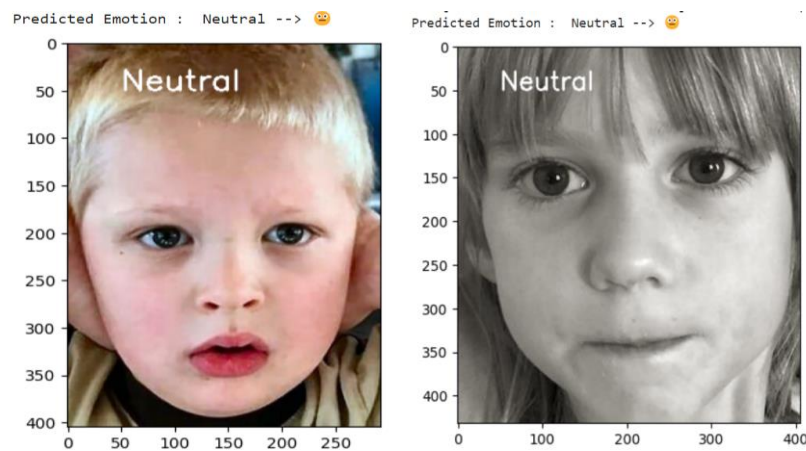


Figure 9. CNN model applied to autistic's images

4. CONCLUSION AND FUTUR WORK

In this study, two deep learning techniques were applied: the CNN model, which has been used on a variety of topics and particularly on facial emotion recognition, has produced very interesting results. The second technique is the ViT, which has been used more recently on facial emotion recognition and image analysis. By examining the images manually, the CNN method assisted in determining the emotions of youngsters with autism. This is the first stage in our work to better understand the autistic person's emotions, provide the right aid, particularly when he is upset or furious, assist him and avoid emergency situations by alerting his parents, assistants. In order to make the analysis and intervention more effective, we look forward to developing the chatbot that offers a method of communication with the autistic child and analyzing his mental state based on real-time analysis of videos rather than just pictures. Also, for better detection of emotions in real-time and engaging with children in future work, it can be advantageous to combine sentiment analysis from speech with facial emotion recognition.

REFERENCES




- [1] F. E. El Rhatassi, B. El Ghali, and N. Daoudi, "Improving health care services via personalized medicine," in *International Conference on Big Data and Internet of Things*, 2023, pp. 435–449, doi: 10.1007/978-3-031-28387-1_37.
- [2] E. Marriott, J. Stacey, O. M. Hewitt, and N. E. Verkuijl, "Parenting an autistic child: experiences of parents with significant autistic traits," *Journal of Autism and Developmental Disorders*, vol. 52, no. 7, pp. 3182–3193, Jul. 2022, doi: 10.1007/s10803-021-05182-7.
- [3] I. Chaidi and A. Drigas, "Autism, expression, and understanding of emotions: literature review," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, no. 02, Feb. 2020, doi: 10.3991/ijoe.v16i02.11991.
- [4] O. Mujtaba Khandy and S. Davdandipour, "Analysis of machine learning algorithms for character recognition: a case study on handwritten digit recognition," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 21, no. 1, pp. 574–581, Jan. 2021, doi: 10.11591/ijeecs.v21.i1.pp574-581.
- [5] O. Higgins, B. L. Short, S. K. Chalup, and R. L. Wilson, "Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: an integrative review," *International Journal of Mental Health Nursing*, vol. 32, no. 4, pp. 966–978, Aug. 2023, doi: 10.1111/inm.13114.
- [6] G. Meena, K. K. Mohbey, S. Kumar, R. K. Chawda, and S. V. Gaikwad, "Image-based sentiment analysis using InceptionV3 transfer learning approach," *SN Computer Science*, vol. 4, no. 3, Mar. 2023, doi: 10.1007/s42979-023-01695-3.
- [7] G. Meena, K. K. Mohbey, and S. Kumar, "Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach," *International Journal of Information Management Data Insights*, vol. 3, no. 1, Apr. 2023, doi: 10.1016/j.ijime.2023.100174.
- [8] K. K. Mohbey, G. Meena, S. Kumar, and K. Lokesh, "A CNN-LSTM-based hybrid deep learning approach for sentiment analysis on Monkeypox tweets," *New Generation Computing*, vol. 42, no. 1, pp. 89–107, Mar. 2024, doi: 10.1007/s00354-023-00227-0.
- [9] G. Meena and K. K. Mohbey, "Sentiment analysis on images using different transfer learning models," *Procedia Computer Science*, vol. 218, pp. 1640–1649, 2023, doi: 10.1016/j.procs.2023.01.142.
- [10] F. M. Talaat, "Real-time facial emotion recognition system among children with autism based on deep learning and IoT," *Neural Computing and Applications*, vol. 35, no. 17, pp. 12717–12728, Jun. 2023, doi: 10.1007/s00521-023-08372-9.
- [11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, Jun. 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.
- [12] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: a database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, Jan. 2019, doi: 10.1109/TAFFC.2017.2740923.
- [13] S. Li and W. Deng, "A deeper look at facial expression dataset bias," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 881–893, Apr. 2022, doi: 10.1109/TAFFC.2020.2973158.
- [14] Y. Khairuddin and Z. Chen, "Facial emotion recognition: state of the art performance on FER2013," *arXiv preprint arXiv:2105.03588*, 2021.
- [15] G. Meena, K. K. Mohbey, A. Indian, M. Z. Khan, and S. Kumar, "Identifying emotions from facial expressions using a deep convolutional neural network-based approach," *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 15711–15732, Jul. 2023, doi: 10.1007/s11042-023-16174-3.
- [16] I. J. Goodfellow *et al.*, "Challenges in representation learning: a report on three machine learning contests," in *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, 2013, pp. 117–124, doi: 10.1007/978-3-642-42051-1_16.
- [17] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning social relation traits from face images," *Proceedings of the IEEE international conference on computer vision*, pp. 3631–3639, 2015, doi: 10.1109/ICCV.2015.414.
- [18] R. Pecoraro, V. Basile, and V. Bono, "Local multi-head channel self-attention for facial expression recognition," *Information*, vol. 13, no. 9, Sep. 2022, doi: 10.3390/info13090419.
- [19] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019, doi: 10.1109/ACCESS.2019.2917266.
- [20] T. Wolf *et al.*, "Transformers: state-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45, doi: 10.18653/v1/2020.emnlp-demos.6.
- [21] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [22] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER: facial emotion recognition with vision transformers," *Applied System Innovation*, vol. 5, no. 4, Aug. 2022, doi: 10.3390/asi5040080.
- [23] F. E. El Rhatassi, B. El Ghali, and N. Daoudi, "A chatbot's architecture for customized services for developmental and mental health disorders: autism," in *International Conference on Digital Technologies and Applications*, 2023, pp. 134–141, doi: 10.1007/978-3-031-29860-8_14.
- [24] A. Vijayan, S. Janmasree, C. Keerthana, and L. Baby Syla, "A framework for intelligent learning assistant platform based on cognitive computing for children with autism spectrum disorder," in *2018 International CET Conference on Control, Communication, and Computing (IC4)*, Jul. 2018, pp. 361–365, doi: 10.1109/CETIC4.2018.8530940.
- [25] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.
- [26] S.-W. Hwang and J. Sugiyama, "Computer vision-based wood identification and its expansion and contribution potentials in wood science: a review," *Plant Methods*, vol. 17, no. 1, Apr. 2021, doi: 10.1186/s13007-021-00746-1.
- [27] A. Vaswani *et al.*, "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, vol. 30, pp. 1–11.
- [28] X. Yao, T. Song, J. Zeng, and Y. Xie, "Rotation invariant convolutional neural network based on orientation pooling and covariance pooling," *The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)*, pp. 433–443, 2022, doi: 10.1007/978-981-16-6963-7_40.
- [29] J. Wu, Y. Jiang, S. Bai, W. Zhang, and X. Bai, "SeqFormer: sequential transformer for video instance segmentation," in *European Conference on Computer Vision*, 2022, pp. 553–569, doi: 10.1007/978-3-031-19815-1_32.
- [30] G. del Castillo Torres, M. F. Roig-Maimó, M. Mascaró-Oliver, E. Amengual-Alcover, and R. Mas-Sansó, "Understanding how CNNs recognize facial expressions: a case study with LIME and CEM," *Sensors*, vol. 23, no. 1, Dec. 2022, doi: 10.3390/s23010131.
- [31] Y. Liu, H. Pu, and D.-W. Sun, "Efficient extraction of deep image features using convolutional neural network (CNN) for

applications in detecting and analysing complex food matrices,” *Trends in Food Science & Technology*, vol. 113, pp. 193–204, Jul. 2021, doi: 10.1016/j.tifs.2021.04.042.




- [32] I. Khan, “Autistic children facial dataset,” *Kaggle*. <https://www.kaggle.com/datasets/imrankhan77/autistic-children-facial-data-set> (accessed Mar. 15, 2022).
- [33] F. M. Talaat, “Autistic Children Emotions-Dr. Fatma M. Talaat,” *Kaggle*. <https://www.kaggle.com/datasets/fatmamtalaat/autistic-children-emotions-dr-fatma-m-talaat/data> (accessed Feb. 16, 2023).

BIOGRAPHIES OF AUTHORS






Fatima Ezzahrae El Rhatassi    achieved her high school diploma in mathematics, followed by completing preparatory classes in mathematics science in 2014 in Morocco. Earned a degree in data engineering from the School of Information Sciences in Rabat in 2017. Commenced her doctoral studies in January 2022 and currently holds a position as a senior data engineer while pursuing her Ph.D. Her research focuses on leveraging artificial intelligence in personalized medicine, analyzing and consuming information, as well as big data and deep learning. Reachable via email at fatima-ezzahrae.el-rhatassi@esi.ac.ma.



Btihal El Ghali    in 2006, was awarded a French baccalaureate with a focus on mathematics, and subsequently pursued higher education by obtaining a bachelor's degree in mathematics and computer science in 2009 from Mohammed V University's Faculty of Science in Rabat, Morocco. Continued her studies at the same institution, where she acquired a master's degree in applied informatics and telecommunication in 2011 and finalized her doctoral studies in July 2016. Presently, she holds a position as an assistant professor at The Information Science School in Rabat, where her research endeavors encompass areas such as artificial intelligence, personalized medicine, information retrieval, and big data. Reachable at bel-ghali@esi.ac.ma.



Najima Daoudi    holds the position of full professor at the school of information sciences, located on Avenue Ibsina B.P. 765 Agdal, in Rabat, Morocco. With an engineering degree from the National School for Computer Science and Systems Analysis and a PhD in computer science from ENSIAS, her research spans ontologies, artificial intelligence, natural language processing, machine learning, MLOps, and data visualization. For correspondence, is reachable at ndaoudi@esi.ac.ma.