IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

FACIAL EXPRESSION FOR PAIN IDENTIFICATION WITH DEEP LEARNING METHODS

S.Anitha[1], Mrs.P.J Mercy[2]

Department of Computer Applications, Sarah Tucker College, Thirunelveli-7.

Abstract: Pain is a strong symptom of diseases. Being an involuntary unpleasant feeling, it can be considered as a reliable indicator of health issues. Pain has always been expressed verbally, but in some cases, traditional patient self-reporting is not efficient. On one side, there are patients who have neurological disorders and cannot express themselves accurately, as well as patients who suddenly lose consciousness due to an abrupt faintness. On another side, medical staff working in crowded hospitals need to focus on emergencies and would opt for the automation of the task of looking after hospitalized patients during their entire stay, in order to notice any pain related emergency. These issues can be tackled with deep learning. Knowing that pain is generally followed by spontaneous facial behaviors, facial expressions can be used as a substitute to verbal reporting, to express pain. That is, with the help of image processing techniques, an automatic pain assessment system can be implemented to analyze facial expressions and detect existing pain. In this project, a convolutional neural network model was built and trained to detect pain though patients' facial expressions, using the UNBC-McMaster Shoulder Pain dataset. First, faces were detected from images using the Haarcascade Frontal Face Detector, provided by OpenCV, and preprocessed through gray scaling, histogram equalization, face detection, image cropping, mean filtering and normalization. Next, preprocessed images were fed into a CNN model which was built based on a modified version of the VGG16 architecture. The model was finally evaluated and fine-tuned in a continuous way based on its accuracy.

I. INTRODUCTION

Humans move their facial muscles, either spontaneously or purposefully, to convey a certain emotional state (e.g., sadness, happiness, fear, disgust, pain) in a nonverbal way. These facial moves are called facial expressions. Facial expressions vary between different species and humans as well; they can be affected by a person's age, gender, psychological state, personality and social situations. Moreover, they can either be innate or acquired through the influence of someone else's. Humans have the ability to discern hidden feelings and fake emotions in some cases, especially when these are expressed by someone with whom they have strong relationships. However, the automation of such tasks is very laborious and challenging.

Facial expressions can be regarded as an effective alternative to verbal communication. For instance, paralyzed people can communicate through eye contact and eye movements. Therefore, facial expressions are very important and worthy to be interpreted by machines, and one of the applications in which they are involved is the detection of pain.

Pain is an unpleasant feeling which is triggered by an anomaly in the body. This anomaly can either be medical (e.g., an injury), or emotional (e.g., stress and depression which can cause terrible headaches). When nerves detect tissue damage or irritation, they send information through the spinal cord to the brain, thus causing humans to react to that anomaly. Pain is either expressed verbally or physically, though facial expressions.

Pain can vary from being slightly annoying to debilitating. Regardless of its intensity, it gives a strong and reliable message that something within the body is malfunctioning and needs to be cured. Additionally, it can affect a person's behaviour, memory, concentration and intellectual capacities. Hence, it should never be neglected and needs to be taken seriously and treated promptly.

Computer Vision

Computer vision is one of the most prominent and most complex research areas nowadays due to the fact that it tries to reproduce the vision task, which is effortless and automatic for humans and many animals but very complex and computationally expensive for computers.

Computer vision encompasses a wide range of applications. From reproducing human visual abilities to creating new visual capabilities, Computer vision applications vary between giving the ability to computers to recognize faces and classify objects into different categories, helping self-driving cars to identify pedestrians, road obstacles and traffic lights, and in a more complex way, synthesizing and restoring defected images and recognizing sound waves from discernible vibrations in videos.

Digital Image Processing

Image processing involves a number of mathematical operations applied to images in order to extract relevant information to perform a certain task. A grayscale image is a matrix of pixels which are stored as 8-bit integers ranging from 0 (Black) to 255 (White). As to RGB images, three matrices (or a tensor), representing the primary colours of light (Red, Green and Blue), are superimposed on each other to represent different colours and colour nuances. Images can be represented in other color spaces, such as the CMYK, which combines the primary colours of pigment (Cyan, Magenta, Yellow and Key or Black).

Deep learning

Knowing why CNNs are the go-to solution in computer vision requires some knowledge about deep learning. Machine learning is an application of artificial intelligence which is being used in any sector one can think of, whether to make predictions, prescriptions or to describe certain behaviors or patterns in data. In simple words, it is the application through which machines learn and improve by experience using a set of examples characterized by predefined features, such as the location, area and number of rooms of a flat.

In some cases, input data is in a raw and unstructured form, which means that its features need to be extracted first before it can be processed. For instance, a picture is a set of pixels, which initially do not have any specific feature. In such a case, deep learning is necessary, since it takes into account feature extraction as well.

Supervised learning:

The term supervised implies that the machine learning system is being guided by an instructor. In this type of learning, each example is labeled. Input features are linked to an output feature, also called a target feature. This means that the model knows what value or class is assigned to which example, and tries to find a relationship between inputs and outputs to predict future outputs for future unseen data. If these outputs are categorical, we are performing classification (e.g., classifying a transaction as fraudulent or genuine). Else, if the output is a numerical real value, we are performing regression (e.g., predicting the price of a flat).

Unsupervised learning:

In this type of learning, outputs are unknown, i.e, the machine learning model has no guidance or assistance. As a consequence, it tries to find similarities between examples and puts them in different clusters, which are similar internally but different externally. This task is called clustering, and it is the most common unsupervised learning task

II. LITERATURE SURVEY

In[1], Facial expressions are signals that have received attention from researchers for many applications, such as face recognition in the field of biometrics.

In[2], Deep learning is used directly to estimate the pain from face expressions. One of the distinct approaches is to estimate the pain from the self-reported visual analog scale (VAS) pain levels to understand the individual differences [1]. Their method includes two learning stages. The first one, which is performed by learning the recurrent NNs (RNNs), estimates the Prkachin and Solomon pain intensity (PSPI) levels from face images. Then, personalized hidden conditional random fields (HCRFs) used the previous output to estimate the VAS for each person. By making comparisons with non-personalized approaches, this approach achieved high performance, and the score for a single-sequence test is the best.

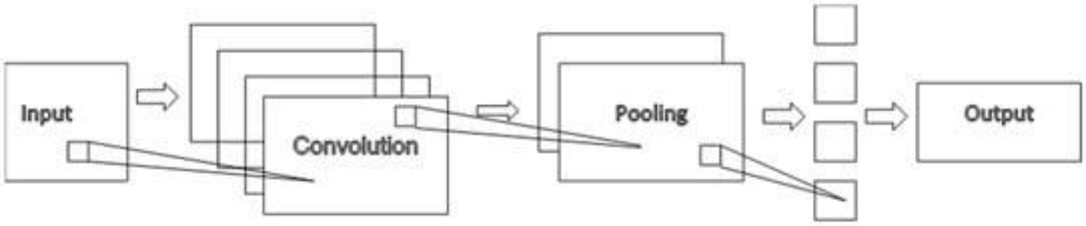
In[3], Deep learning has been mainly used to extract the important features, as was recently done by [2] for pain detection based on facial expressions. Their approach is based on three steps: First, convolutional neural networks (CNNs) are used to extract the features from VGG_Faces. After that, the result of the feature map is used to train the LSTM. This is a type of RNN used to find the binary pain estimation (pain, no pain). They provided a summary of previous works done on the popular data set of pain detection using faces. This data set is called a UNBC–McMaster database and has 200 video sequences from 25 patients who suffered from shoulder pain. Their experiments on this data set showed that their approach outperforms all previous works with an area under the curve (AUC) performance of 93.3%. Their model can also be generalized for application to other facial emotion recognition. This ability was realized when applying their model to the Cohn Kanade + facial expression database and resulted in a competitive score (AUC = 97.2%).

In[4], In the same manner, in 2017, Egede, Valstar and Martinez [3] proposed a pain-estimation model that combined learned features obtained from deep learning and other handcrafted features. Their idea comes from the hardness to obtain a data set for pain estimation in a large area in order to work well with deep learning. Therefore, they extracted handcrafted features directly from the face image and used a CNN to learn some features. Their features include appearance, shape and dynamics information. Finally, they classified the pain level using the linear regression model on the combined features and individual. Their results outperformed the state-of-the-art methods in terms of the root mean square error (RMSE) of 0.99 and Pearson correlation (CORR) of 0.67. A limitation of this approach—and all face-based approaches—is that they consider only the front of the face without capturing several combinations of indicators, such as audio, body movements and physiological signals.

In[5], Another solution that aims to deal with small data sets and deep learning was proposed in 2017 by Wang et al. [4]. They fine-tuned a small pain data set using a face verification network that is trained by the WebFace dataset, which has 500,000 face images. Then, they fitted a problem as a regression problem by applying a regression loss regularized with the center loss. Their performance was evaluated based on new proposed metrics to avoid the use of imbalance data. Based on the results, this method achieved a high performance compared with the state-of- art methods using both weighted metrics (mean absolute error (MAE): 0.389, (mean squared error) MSE: 0.804, (Pearson's correlation coefficient) PCC: 0.651) and new proposed metrics (weighted MAE 0.991, weighted MSE 1.720). However, pain is temporal and involves subjective information, and no such information and stimulus knowledge are used in this method, which requires further investigation.

III. METHODOLOGY

NNs were derived from the first model on an NN that was invented in 1998. In general, conventional networks perform logistic regression by applying a filter to the input. In addition to the filter size (f), the important parameters required to build a deep CNN are the stride (s) and padding (p). CNNs have many types of layers, namely conventional, pooling and fully connected layers.



The CNN has an advantage of generalization compared with MLP. In addition, it has fewer parameters than the fully connected layers in MLP.

As we can see from the literature, CNNs are mostly used for feature extraction. Only two studies have employed them for classification tasks. However, they were used only for pain recognition based on facial expressions without considering physiological signals and speech analysis.

Data Preprocessing

Real-world data is often subject to errors, noise and outliers. Before it can be visualized or processed, it needs to be preprocessed and cleaned. Preprocessing is a crucial phase in data analysis. It allows us to do a data quality assessment and resolve all issues which might affect the performance of our model.

The following preprocessing operations have been used in this capstone project, using the OpenCV library:

Gray scaling, to keep one channel in each image and simplify the convolution operations. Colors are not impactful in this problem. However, they can be of the essence in other problems (e.g., the classification of fruits into different quality classes based on many factors, including their color).

Histogram equalization, to unify and improve the contrast of every image for better edge detection. Consequently, images would neither be too bright nor too dark.

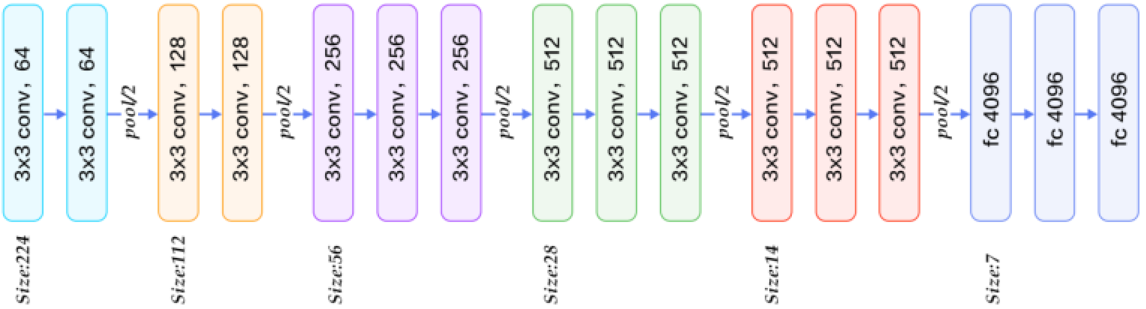
Face detection, using the predefined Haarcascade Frontal Face DetectorImage cropping, to keep the frontal face of the patient only

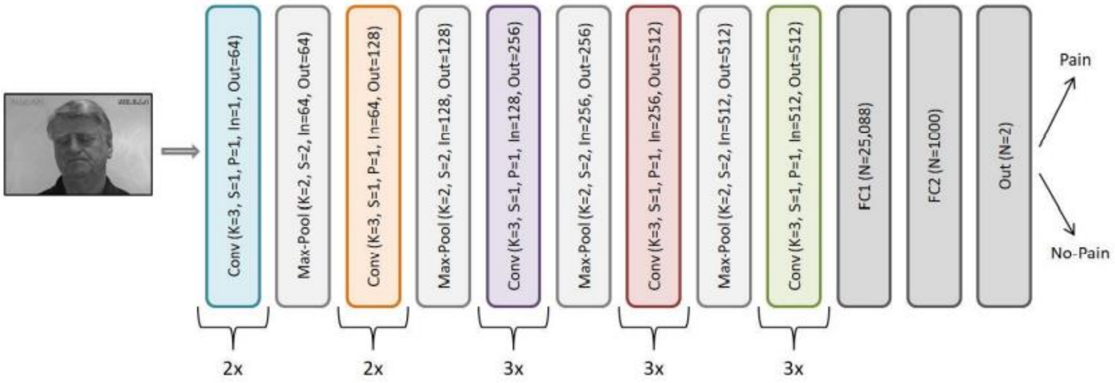
Mean filtering, to eliminate unrepresentative pixels. In this step, every pixel is replaced with the average value of its neighbors, including its own value.

Normalization/Standardization, to keep pixel intensities within the range [-1,1] and have small standardized values.

CNN VGG17

CNNs have been used previously to solve the automatic pain assessment problem, using the UNBC-McMaster Shoulder Pain dataset. Based on existing architectures and after many trials and hyperparameter tunings, the architecture shown in Figure has been tailored for this project, and it is a modified version of the famous VGG16 architecture, shown



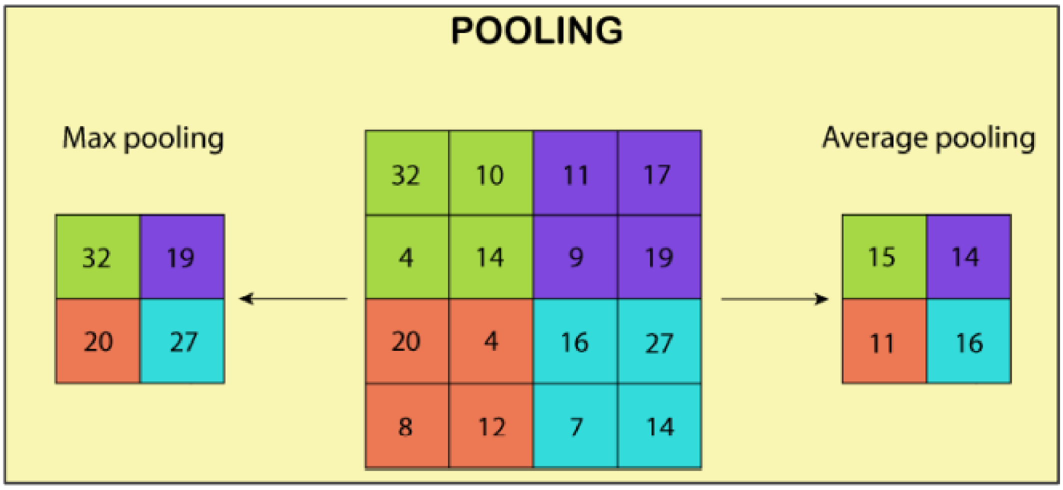

Conv (K=3, S=1, P=1, In=1, Out=64)
Max-Pool (K=2, S=2, In=64, Out=64)
Conv (K=3, S=1, P=1, In=64, Out=128)
Max-Pool (K=2, S=2, In=128, Out=128)
Conv (K=3, S=1, P=1, In=128, Out=256)
Max-Pool (K=2, S=2, In=256, Out=256)
Conv (K=3, S=1, P=1, In=256, Out=512)
Max-Pool (K=2, S=2, In=512, Out=512)
Conv (K=3, S=1, P=1, In=512, Out=512)
FC1 (N=25,088)
FC2 (N=1000)
Out (N=2)
Pain
No-Pain
2x
2x
3x
3x
3x

Convolutional Neural Networks (CNNs) are one type of artificial neural networks which are widely used to perform computer vision tasks because of their ability to process huge data in a more simplified and organized way than ordinary neural networks. The logic behind CNNs is very similar to ordinary neural networks. The major difference is that in ordinary neural networks, each input is represented by a neuron, and each neuron in a layer is connected to the next layer's neurons. This is called a fully connected network. In CNNs, there are groups of neurons, and each group is responsible for processing one area of the image.

CNNs use three types of layers:

1. The convolutional layer: In this type of layers, each group of neurons processes one area of the image, (i.e., a group of pixels from the image) using:

2. A Kernel or Filter, which is a tensor of size (fw x fh x nc), with fw and fh being the width and height of the filter respectively (fw and fh are usually odd), and nc the number of channels in the image. Each channel of the filter is superimposed on the appropriate channel of each region in the image. The values of each filter's channel are multiplied by the targeted region's values in the appropriate channel and summed to get a scalar value. Thus, the dimensions of the image decrease after each convolution, as shown in formulas (3.10) and (3.11) and Figure 3.5. Many filters can be used in one layer, in order to detect as many different features as possible. For instance, in edge detection, using many filters helps the model to detect features with different orientations. Each filter has its own bias. Thus, the number of parameters to learn in each convolutional layer equals the size of one filter multiplied by the number of filters, plus the number of filters, which corresponds to the biases.
3. A stride, which is the number of pixels by which the filter slides between regions.
4. The pooling layer: Since many filters are used in each convolution, the depth of the input image increases. Thus, in order to reduce computational time and complexity, CNNs reduce the convolution outputs' dimensionality, in the pooling phase, by applying an empty filter on each region (as explained in the convolutional layer), which extracts the most important features only. Since a pooling filter is empty, there are no parameters to learn in a pooling layer.

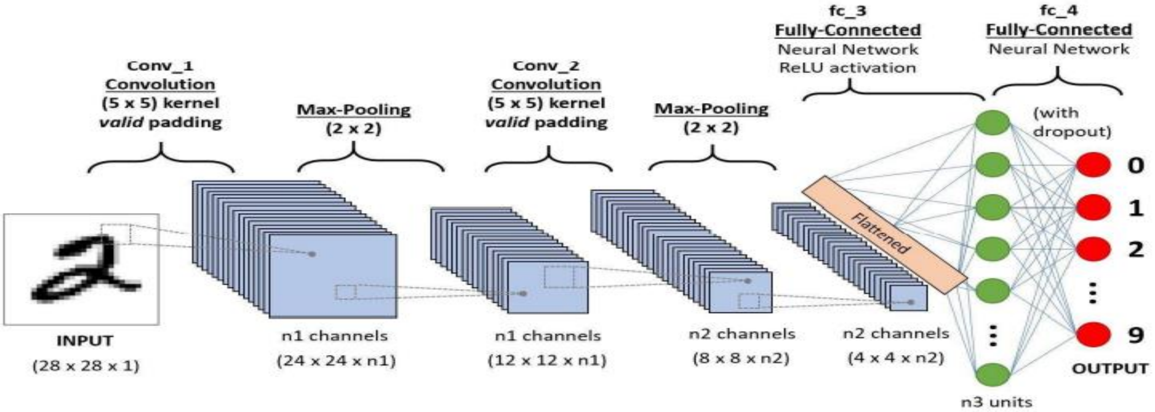


32	10	11	17
4	14	9	19
20	4	16	27
8	12	7	14

32	19
20	27

15	14
11	16

5. The Fully Connected (FC) layer: Once all features are extracted, they are gathered in a tensor of features, called a feature map. This feature map is then flattened as a vector, called a fully connected layer, because each neuron in this vector is fully connected with all neurons in the next vector. This flattened layer becomes the input layer of a simple neural network.
6. A convolutional layer can either be followed by another convolutional layer, a pooling layer or a fully connected layer. Similarly, a pooling layer can either be followed by a convolutional layer or a fully connected layer. But the latter is the beginning of a fully connected network, so it is automatically connected to another fully connected layer.



7. OpenCV is a computer vision and Machine Learning library aimed at performing different operations on images such as face recognition and movingobjects tracking, in addition to more elementary transformations like gray-scaling, morphing, normalization and rotation. OpenCV is widely used by a lot of big companies such as Google, Microsoft and IBM.

IV. EXPERIMENT AND ANALYSIS

VGG19.

The convolutional neural network (CNN) is shown in Figure 1, which includes a convolutional layer, a down sampling layer, and a fully connected layer. Each layer has multiple feature maps, and each feature map has multiple neurons, and the input features are extracted through the convolution filter. The parameter sharing mechanism of the convolutional layer greatly reduces the number of parameters.

The research is based on the VGG19 network to optimize and improve the network. The main structure of VGG19 consists of 5 convolution modules, 3 fully connected layers, and an input layer and output layer. Each convolution layer module is down sampled through the max pool.

The expression of the convolutional layer is as follows:

$$x_j^l = f\left(\sum_{i\in M_J} x_i^{l-1} i * k_{ij}^l + b_j^l\right)$$

In Equation (1), assuming that l − 1 is the input layer or the pooling layer, and the l layer is the convolutional layer, then x_i^l is the j-th feature map of the l convolutional layer; the right side of Equation (1) represents the feature map of the l − 1 layer. Perform convolution operation with the j-th convolution kernel k_{ij}^l of the l layer and sum; b represents the bias; f(·) is the activation function ReLU.

The pooling layer closely follows the convolutional layer and plays the role of scaling dimension a subj The calculation equation is as follows:

$$x_i^l = f\left(\beta_j^l \mathrm{down}\left(x_i^l - 1\right) + b_j^l\right). \qquad (2)$$

In Equation (2), down(·) is the pooling function, which seeks the maximum feature map region of the feature map; β_j^landb_j^l , respectively, represent the weight and bias of pooling.

The input layer size of VGG19 is 224 × 224 × 3, and the convolution module is composed of a stack of convolution layers and pooling layers. The convolution kernel is usually 3×3 with a step size of 1, and the pooling layer is a 2×2 max pool. Using the convolutional layer and the pooling layer to cooperate, on the one hand, the image size is reduced and the amount of model calculation is controlled. On the other hand, the convolution data of the large receptive field is obtained indirectly, and the high-dimensional feature map is obtained. The convolution module is followed by three fully connected layers to obtain the classification information of the feature map, and finally, the softmax layer is used to output the classification results. The structure diagram of the VGG19 network is shown in Figure 2.

The increase in the depth of the convolutional neural network in the VGG16 network and the use of small convolution kernels have a great impact on the final classification Computational and Mathematical Methods in Medicine structure.

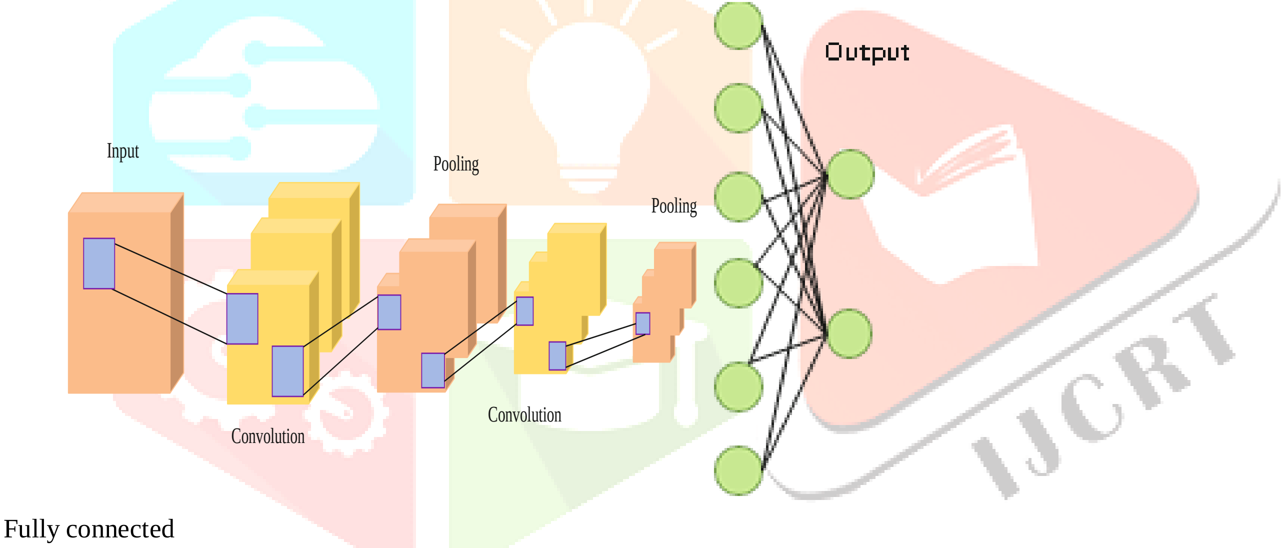


Figure 1: Convolutional neural network

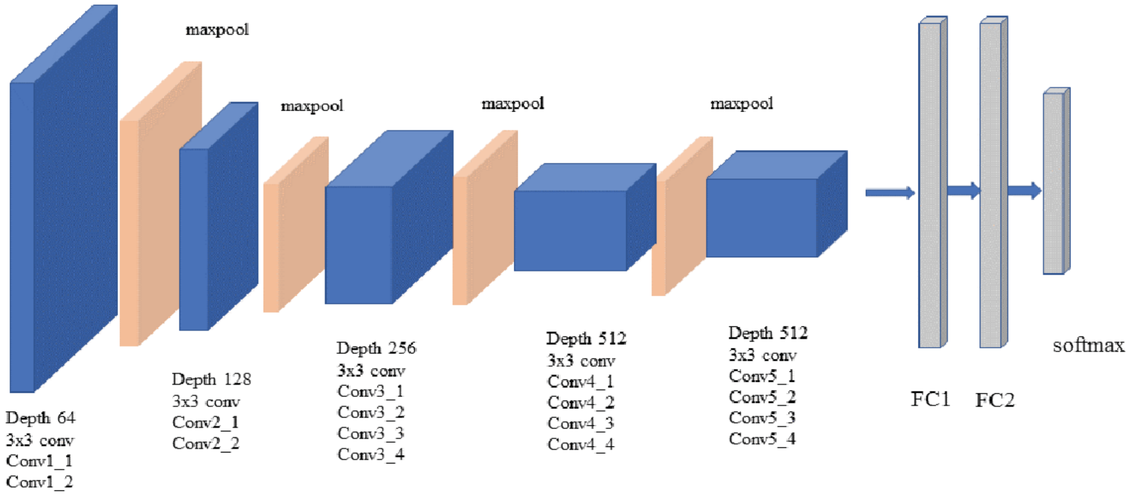


Figure 2: VGG19 structure

and recognition effect of the network. The convolutional layers all use the same 3-size convolution kernel parameters, and the pooling layers all use the same pooling kernel parameters. The combination of multiple 3×3 convolutional layers not only has a small amount of calculation but also obtains the same receptive field of the large convolution kernel at the same time. The deep network structure verifies the conjecture that network performance can be improved by continuously deepening the network structure. But for some data, a too deep network only greatly increases the training time but does not improve the accuracy. The convolution kernel of VGG16 increases from 64 to 512 sequentially and the number of image channels are first reduced to 64 and then increased to 512. However, due to a large amount of image data, this change in the number of channels will cause the data to lose a lot of information. Increasing the time cost of training and the network structure of VGG16 for this research task, while increasing the depth of the network, cannot improve the accuracy of the network.

Convolutional neural networks are mainly composed of convolutional layers, nonlinear units, pooling layers, and fully connected layers. In the class- sification problem, the convolutional layer, the nonlinear unit, and the pooling layer are used as the feature extraction layer to extract features, and the fully connected layer is used as the classification layer for classification. The convolutional layer is the core of the convolutional neural network, and the convolution equation is shown in Equation (3).

$$y(t) = \int_{-\infty}^{\infty} x(P)h(t-P)\,dP = x(t) \times h(t). \qquad (3)$$

The nonlinear unit is the ReLU activation function, and its expression is shown in Equation (4).

The pooling layer is a downsampling operation to reduce the dimensionality of the extracted features while retaining important information about the features.

The VGG19 network is trained on a large data set Ima- geNet. The ImageNet data set is a 1000 classification problem data set, so the classification layer parameters of the VGG16 network are huge. The diagnosis of schizophrenia is a two-class classification problem and does not require a complex classification layer. Therefore, the feature extraction layer of the VGG19 network is retained, the classification layer is redesigned, and the original 3-layer fully connected layer is improved to a 2-layer fully connected layer. We take the features of 3 convolutional layers and 3 pooling layers as an example, and the process of part of the extracted features is shown in Figure 3. Use the ReLU activation function, add a dropout layer to prevent over- fitting, and change the final output classification to two categories. The data can be divided into schizophrenia and nonschizophrenia, and the amount of parameters is reduced so that the network converges faster, and the recognition speed of the data is improved. Figure 4 shows the improved VGG19 schizophrenia classification model.

Transfer Learning. Transfer learning solves the shortcomings of deep learning that requires a large number of sample training models. By training a pre-trained model on a large data set, it is possible to use a small number of data sets to train the model. Fine-tune is a training method that retains the model feature extraction layer and retrains the model classification layer. The pretraining model used is the VGG19 network pre-trained on the ImageNet data set, and the feature extraction layer of the pretraining model is fixed. Retrain the improved classification layer of VGG19 to complete the training of the schizophrenia classification model.

V. CONLIUSION

Computer vision is a topic which is in full effervescence. Being the subject of many research papers nowadays, it keeps evolving and it is progressively becoming part of the 21*st* century's AI applications, automatic pain detection is one of them, as we saw in this project.

Deep learning keeps proving its worth since a couple of years now, especially with huge datasets which need deep processing, and CNNs gained a lot of popularity and success with this kind of data. This project has specifically confirmed this point. We saw that with relatively big images of 224 x 224 pixels = 50,176 inputs, CNN-based models can perform complex and very deep processings effectively and neatly. With tools such as OpenCV, the implementation of very complex architectures can be easy and time-saving.

Healthcare systems are complex systems that contain interactions between different entities: people, process and technology. This study is considered as a starting point for researchers to develop a smart healthcare system. It can provide them with available tools and datasets to build such systems.

This study presents a systematic review of pain-recognition systems that based on deep-learning methods only. Based on the papers reviewed, a new taxonomy of categorization was presented based on the kind of data used. These pain-recognition data were obtained from facial expressions, speech, or physiological signals. Furthermore, this study describes the primary deep-learning methods that were used in review papers. Finally, the main challenges and future direction were discussed.

Deep-learning algorithms have many advantages in healthcare systems. The biggest advantage is their ability to describe the complex problems and non objective measures such as pain. They could extract the features automatically without a fully understanding of health problem from medical experts. In addition, the difficulty of collecting large data from patients could be overcome by using suitable augmentation techniques. Therefore, the intelligent interpretation of problem by deep learning and also increasing the medical test data automatically will enhance the rapid development of smart healthcare systems.

REFERENCES

1. Martinez, D.L.; Rudovic, O.; Picard, R. Personalized Automatic Estimation of Self-Reported Pain Intensity from Facial Expressions. *arXiv* **2017**, arXiv:1706.07154. Available online: **http://arxiv.org/abs/1706.07154** (accessed on 26 July 2018).
2. Rodriguez, P.; Cucurull, G.; Gonzalez, J.; Gonfaus, J.M.; Nasrollahi, K.; Moeslund, T.B.; Roca, F.X.; López, P.R. Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification. *IEEE Trans. Cybern.* **2017**, 1–11. [**Google Scholar**] [**CrossRef**] [**PubMed**]

3. Egede, J.; Valstar, M.; Martinez, B. Fusing Deep Learned and Hand-Crafted Features of Appearance, Shape, and Dynamics for Automatic Pain Estimation. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 689–696. [Google Scholar] [CrossRef]

4. Wang, F.; Xiang, X.; Liu, C.; Tran, T.D.; Reiter, A.; Hager, G.D.; Quon, H.; Cheng, J.; Yuille, A.L. Regularizing face verification nets for pain intensity regression. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 1087–1091. [Google Scholar] [CrossRef]

5. Jaiswal, S.; Egede, J.; Valstar, M. Deep Learned Cumulative Attribute Regression. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 715–722. [Google Scholar] [CrossRef]