National College of Ireland

Facial Emotion Recognition using Deep Convolutional Neural Network

MSc Research Project

MSC in Data Analytics, MSCDAD_JAN21A_I

Sachin Pralhad Langute

Student ID: 19234635

School of Computing

National College of Ireland

Supervisor: Prof. Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:

Student ID: ……19234635……

Programme: MSC in Data Analytics MSCDAD_JAN21A_I **Year:** …2021-22……..

Module: ……Research Project……

Supervisor: ……Prof. Jorge Basilio……
Submission Due Date: ……31st January 2022……

Project Title: …Facial Emotion Recognition using Deep Convolutional Neural Network…

Word Count:……6637…… **Page Count**……16……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Sachin Pralhad Langute……

Date: ……31st January 2022……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Facial Emotion Recognition using Deep Convolutional Neural Network

Sachin Pralhad Langute

19234635

Abstract

The recognition of facial expressions (FER) is critical for social communication. However, existing research has limitations when it comes to addressing facial expression differences related to variations in the demographics such as age, gender, etc. In face-to-face encounters, facial expressions communicate nonverbal information. Since the last three decades, researchers have become increasingly interested in detecting facial expressions automatically, which is very important for Human-Computer interaction. The average person exhibits seven various emotions depending on the scenario, which include anger, sorrow, happiness, surprise, disgust, neutral, and afraid. Every person has their style to show emotions, which cannot be related culturally. Traditional machine learning algorithms can need a sophisticated feature extraction procedure and yield poor results. Artificial Neural Networks (ANN) have been developed to address some of these constraints. The latter produce good results but do not address all the issues such as camera angle, head position, occlusions, and so on. In this research, the author investigates neural network models that are employed in the field of face emotion identification. The author also offers a bilinear pooling-based architecture to build on earlier work's achievements and to give solutions to these reoccurring restrictions. This method vastly increases the performance of designs based on traditional CNNs. This study investigates deep learning strategies for face emotion identification based on Convolutional Neural Networks as well as the VGG16 model. Furthermore, input data is extended by rotation, cropping, and flipping.

Keywords: Facial emotion recognition, Convolutional Neural Network (CNN), VGG16, Deep Learning, Classification, Machine Learning.

1 Introduction

The remembrance of emotions is critical during the production of Augmented Reality (AR) and Artificial Intelligence (AI) in understanding human actions for various physiological and psychological elements in the fields of medical, computer and mobile apps, sports, and other diverse sectors. Emotion identification may be detected using discrete motions such as face motioning, upper body movements such as hand movement, and total body skeletal movement. The goal of this study is to provide a unimodal recognition system that will provide us with the emotional condition of a real-time human utilizing various gesture representations. This will be advantageous in terms of learning more about human communication behaviors and human-robot operations. Convolutional Neural Networks will be used extensively in this machine learning modeling. Emotion recognition allows users to engage with machines and their surroundings by employing various physical movements to gain a deeper grasp of the

human emotional state. This behavioral condition can aid in a wide range of research projects in a variety of fields. These fields include autonomous assisted driving in automobiles, assisting youngsters with autism spectrum disorder, emotional-intelligence systems such as the Mascot robot system, psychological aid, human-computer interaction, and many others. This emotional detection system simplifies human life by functioning in such a manner that it provides individuals with automated support based on their moods. For example, an ER system may be utilized in a home to adjust the lighting and play music based on the user's psychological condition. Furthermore, it may provide temperature support by measuring human body temperature, which can help to calm people down and make their house a more relaxing environment.

As deep learning allows the usage of various deep neural layers for distinct learning methods, it results in more fame which marks the rise in the applications in the field of computer vision using these practices. CNN gives more precised accuracy and performance as it associates the extracted features with the classified objects in one single learning architecture. Convolutional Neural Network-based learning algorithms are used for segmentation, classification, and reconstruction problems(M. Aza et al. 2020). However, with the introduction of deep learning in the last decade, FER technology has reached exceptional accuracy in identifying emotions from face photos under real situations, outperforming human performance. This has enabled the creation of ground-breaking applications in robotics, health care, automated driving, along with a variety of additional human-computer interaction systems (Yang et al. 2020). Current Human-Machine Connection (HMI) technologies lack the area of socially interacting with people making it difficult to understand emotions. To overcome this challenge and understand human emotions more closely, HMI has made use of facial expressions which has helped a lot to evolve and comprehend social communication (A. Mollahosseini. 2018).

Interaction Human-Machine (IHM) research mostly focused on developing strategies depending upon utilization of mouse, screen, and keyboard. Nowadays, there are no obstructions for the user enabling the use of numerous IO devices, and also using other recognizable techniques such as hand gestures, fingerprints, or face detection. Due to this, there is not much difference left between the real and the artificial world. To achieve this significant evolution, there is a need to intervene in an individual's activity and look into them very carefully which requires computer vision technologies(A. Fathallah et al. 2017).

The goal of this research project is to recognize emotions so correctly that when a real-time image is supplied to the trained model, it should give us the right answer with no risk of error, just like it did on the train/test dataset. It employs a Convolutional Neural Network for the novel capacity of automatically identifying human emotions to assist persons who have partial or complete communication sense impairment. This will enable the intended audience to recognize and trigger a real-time emotional response, as well as offer individuals an emotional stimulation based on Ekman’s classifications of emotions: anger, disgust, fear, happy, sad, surprise, and neutral. Few scholars, however, believe that there are more complex and affective states of emotions that are difficult to categorize in a single discrete class since these emotions may be portrayed in an n number of facial expressions and bodily gestures that must be intervened since these complicated emotions might assist us in comprehending the intellectual mental state (Baron-Cohen et al. 2003). These distinct emotions, which might include shame, depression, agreement, pondering, curiosity, concentration, and so on, can be more instructive about an individual's mental state since they occur more frequently in our everyday lives than fundamental emotions. There are a few other disadvantages to the system, such as picture quality owing to poor lighting and low resolution, distance from where the image is taken, and image complexity due to brightness change, which may disrupt the original image specification (Cootes et al. 1995). According to a study in psychology, there are three critical methods to

emotion modeling that we can distinguish: categorical, dimensional, and appraisal-based approaches (Grandjean et al. 2008).
The primary goal of this study is to analyze gestures and predict human emotions. Thus, the study question for this work might be stated as:

"How correctly can we anticipate human emotions based on facial expressions portrayed in the images?"

Motivation comes from accurately predicting these pictures despite visual complexity, as well as from including an external image into the model and testing accuracy after the model has been trained. The researcher will also attempt to train the model using a collection of physical motions, as these gestures can be employed when facial expressions are not obvious. The Feedforward Deep Convolutional Neural Network will be used in the study in a variety of ways. When we compare with the monomodal, Bi-modal, and multi-modal depending on the accuracy, the multi-model system delivers magnificent outcomes with ultimate high accuracy. This is because mono-models will only include one component of human gesture (for example, only facial expression), however, if other aspects are included in the model, it will be more attentive based on other bodily motions and provide results more pro-actively (Santosh Kumar et al. 2019).

2 Related Work

Many scientific investigations have been stated when it comes to emotion identification since it outshines current advances in every conceivable way. Modern self-driving vehicles use these technologies to automatically switch to autopilot mode when it predicts the driver's mental state or when he or she is driving inattentively (Agrawal et al. 2013). Because of online classes, a relatively current example of an emotion detection system may be observed in the teaching sector. These online platforms may be used to monitor learning activities and assess instructional outcomes. It is possible to execute how these lessons are carried out using the emotions recognition system based on the learning rates and by examining the emotional states of the students (Liang 2019). Vu et al. (2011) presented a bimodal based on dual unimodal as the basis to include informative recognition from speech and gesture that are blended utilizing weights criterion and majority vote procedures. These criteria would assist us in developing a better classifier, allowing us to detect communication circumstances more effectively. In this model, fifty Japanese vocabulary and eight types of gestures from five persons were employed, and an 85.39 percent accuracy rate for emotion identification was attained utilizing RT middleware. Another study claimed that gestures might play an important role in communicating emotional information. This article exclusively employs gestural communication for emotional detection, and it achieves an accuracy of 94.4 percent utilizing fuzzy sets and a training set of only one gesture per feeling (Kar et al. 2013). The system estimates the gestural signals to determine the class of the indicated emotion, implying that emotion recognition results in lower-cost efficiency and higher accuracy with less complexity. They utilized gestures from the six trajectories with the most mobility since the progressing regions supply more data on a person's emotional state, and these datasets are gender agnostic, giving us good accuracy with any external picture input to the algorithm. According to Mishra et al. (2017)'s research, facial expressions account for 93 percent of human communication. Thus, they employed the fundamental seven emotions to demonstrate the intensity for a human-computer communication system and achieved an accuracy of 63.03 percent utilizing the support vector machine and Convolutional neural networks. This study strategy and the

algorithms yield noteworthy results for further investigation in the field of computer-based emotion recognition systems.

The system assesses gestural signals to identify the emotion class, meaning that emotion identification results in lower-cost efficiency and greater accuracy with less complexity. They used gestures from the six most mobile trajectories because they provide more data on a person's emotional state, and these datasets are gender agnostic, providing us high accuracy with any external picture input to the system. Facial expressions make for 93 percent of human communication, according to Mishra et al. (2017)'s research. As a result, they used the basic seven emotions to display the intensity for a human-computer communication system and obtained an accuracy of 63.03 percent using the support vector machine and Convolutional neural networks. This research technique and the algorithms produce notable outcomes for future research in the field of computer-based emotion recognition systems. The data were merged at both the Bayesian feature level and the decision level, and the experiment showed that when data fusion is performed at the feature level, the results are significantly better than when fusion is performed at the decision level. Various psychological research suggests that collecting physical movements and gestures reveal important information on emotion identification, however, few trials use these gestural motions. When developing an emotion recognition system, most theorists rely only on facial expressions and voice. According to recent research, there are many more psychologically sophisticated human emotions, which will improve the overall accuracy of the sentiment and emotion identification system. The authors employed the hashing approach to extract the essential points from the video portrayal and then convolutional Long Short-Term Memory (LSTM) networks to manipulate the series information in this article (Son Thai Ly et al. 2018).

In the realm of Human-Machine Interaction, Emotion recognition has gained a lot of popularity due to the promise of applications. Nonverbal gestures play an important part in transmitting feedback and emotional condition to the user. For such emotion-conveying devices, a plethora of sensors have been used. These sensors can assist robots in learning about the social intelligence of the human species. Facial expression data is collected and put into algorithms for assessing a person's emotional state based on their facial expressions. Furthermore, these datasets are diversely examined in Augmented Reality utilizing a mixed reality device known as Microsoft HoloLens. The study evaluated the outcomes of ER using Microsoft HoloLens and a basic camera (Mehta et al. 2018). It was discovered that photographs recorded with a Microsoft HoloLens have higher accuracy than images acquired with a conventional camera since the webcam images were confused with other emotions such as melancholy. Because of advances in computer science, people may now communicate with machines in previously imagined ways, going beyond manual hardware communication, by employing current ideas such as gestures, voice, and force-feedback networks. However, these advancements miss the most vital component of communication, which is the emotional touch. Several programs are favored to comprehend human emotions to communicate more effectively. Some advances were made in one of the studies, which included the use of facial expressions, voice, and psychological cues in a multimodal system (Sebe et al. 2015). These modalities are treated and trained individually before being merged to demonstrate the obstacles that these multimodal face. Furthermore, the authors describe the numerous limitations and challenges encountered for emotion detection systems, as well as a description of how unlabeled data may be added to the system and used to improve the model's efficiency. One of the theorists suggested the concept of emotional recognition in the teaching domain to better understand students' learning behaviors in the context of modern online education (Wang 2021). The author emphasized the limits of current learning systems that can be used to assess teaching skills by evaluating learning habits. These constraints may be identified by examining the learning behavior utilizing pictures for emotion recognition. The picture emotion identification was first grasped

with the specifics of the online learning behaviors, and then the essential points from the face photos were extracted using the enhanced Local Binary Pattern (LBP) technique and wavelet processing. Later, based on the assessment, the author created an online learning behavior composition and then provided a technique for emotion identification via facial movements based on the attention mechanism. This approach is useful for online learning behaviors based on visual emotion recognition. This model attained an overall accuracy of 80.04 percent by utilizing the convolutional neural network and the attention mechanism. Human-Computer Interaction has recently gained popularity because it may be used to comprehend and communicate with machines based on an individual's emotional state. This research may be used in a variety of fields, including medicine, education, and so on. Physiological signals and neuro-imaging breakthroughs, in addition to facial expressions and voice, can be used to achieve a superior emotion identification system. To attain a greater classification rate, many researchers have adopted a user-dependent emotion recognition system. Larger data samples, as well as sophisticated signal processing techniques, are necessary to achieve this, which will also improve the user-independent classification rate. Jerritta et al. (2011) recommended that physiological signals be used to create an efficient emotion identification system. In the case study, many notions of emotions and actions are examined, as well as the obstacles faced by the emotion identification system using physiological information. The user-dependent emotion recognition system has a 95 percent accuracy rate.

There has been research that has employed many models to get a more accurate emotion identification system using only facial expressions. The Haar cascade, an Adaboost classifier for categorizing fundamental emotions like anger, pleased, disgust, startled, and neutral, and an Active shape model for feature extraction utilizing the 26 face locations are among these models. This implementation is done in real-time on a Raspberry Pi II, with a 94 percent overall accuracy (Suchitra et al. 2016). When this Raspberry Pi II is connected to a mobile or computer-based auto-bot, it can recognize real-time dynamic emotions on a variety of social platforms where emotional activity is required. Rather than testing with the static emotional state of the face, the authors of another study employed dynamic movement abilities such as amplitude, fluidity, and speed of movement (Castellano 2007). Both the time series model and the dynamic motion expressive model are used to train the model. Using a classification technique, the author ended the study by evaluating emotion recognition rates for both models. The modalities of gestures have a considerable impact on the emotion detection system's categorization. On our database, research was recommended that would employ 3-dimensional skeletal data and would apply a posture evaluation basis approach for the extraction of 3D skeletal coordinates (Shi et al. 2021). Early ideas were utilized to link the user's actions to a graphical neural network, which was then used to represent the joints with spatial connections. For skeleton-based emotion identification, a self-attentive enhanced spatial-temporal graph convolutional network is used. The skeletal structure is represented graphically as a static point, and the self-attention model dynamically creates multiple connections between the joints, contributing more information to the system. With an accuracy of 82 percent, this model beats the other models, indicating that the skeletal-based technique used in this work may be incorporated in a multimodal system for maximum emotion identification accuracy. We are now going to look at both the monomodal and multimodal approaches in a single research paper, where the emotions recognition system is first used with the help of bodily gestures, and then the results are compared with the other modality, which is the automatic emotional body gesture recognition (Corneanu et al. 2018). To begin, we use both the dynamic and static position estimation methods to detect people, followed by the use of emotion expressive photos. In another study, the authors used upper body gestures such as head and hand movements by individuals to develop a model for expression analysis of bodily gestures and motion cues for

basic emotions such as anger, relief, sadness, joy, and so on using a scenario-based methodology (Glowinski et al. 2008).

3 Research Methodology

In a Convolutional Neural Network(CNN), the algorithm used is often allowed to learn more composite functions by adding more convolutional layers which results in a better performance of the model. Meanwhile, these increased neurons can also lead to complexities like additional calculative requirements and overfitting of data. In a study by Arora et al. (2013) about the deep sparse networks, the drawbacks of the increased layers can be clearly seen in both the theoretical and biological aspects of the research. The recent CPUs and GPUs lack consistency in computing the problems in sparse networks. To overcome this challenge, Inception layers are used along with the deep convolutional layers in the model used to complex functionalities in sparse networks(O. Russakovsky et al. 2014).
Significant results were obtained by the introduction of the Inception layer in a study carried out by Y. Sun et al. (2015) and C. Szegedy et al. (2014) which was considered to be motivating the upcoming Facial Expression Recognition studies by using the designated techniques. The Inception layers make a significant improvement theoretically depending on the increased size of the network, but this also results in an improvement in feature identification for the local features. As the smaller number of neurons are responsible for local feature identifications, these increased neural layers are accountable for identifying features globally. Each individual can perceive emotions just by looking at the local facial features like lips and eyes movements(E. Bal et al. 2010). This study can be related to children with autism frequently struggling to discriminate between emotions unless they are reminded to look at the same local characteristics (E. Bal et al. 2010). We may expect considerable increases in local feature performance by employing the Inception layer in the algorithm and increasing the number of neural layers study presented by Lin et al. (M. Lin et al. 2013), which can be used to explain the increased performance in the Facial Expression Recognition system.
The other benefit of having more neurons in an algorithm can be seen as reducing the overfitting of trained data. This is achieved as the overall performance of the global pooling is enhanced. This significant improvement in overfitting permits the user to increase the number of convolutional layers in a network and increase the overall performance and accuracy.
The goal of the suggested technique is to construct an emotional detection system based on an open-source Kaggle dataset, where we first train our model using a deep convolutional neural network to recognize fundamental emotions based on facial expressions in photos.

3.1 Dataset

There is a "pixel" field made up of 48*48 face photos. It is saved as a one-dimensional string that has been flattened. Some of these photos are regarded as noisy. For example, funny pictures, images with obscured facial expressions, and images that are entirely dark. None of these photos were eliminated from training for comparison with the earlier study. In the beginning, the author will pass our dataset of a total of 28,709 face photos in model development, with each image being in the format 48*48 pixels grayscale image and the test images consisting of just the pixel value of seven unique emotions (Angry, disgust, fear, happy, sad, surprise, neutral).

Figure 1: Sample images from the dataset.

The following is a link to the dataset:

Facial Expression Recognition Challenge Dataset | Kaggle

For the feature extraction step, we'll center the face in each image such that each image takes up the same amount of area. Then the cleaned and formatted data is populated to the CNN model. To process each image in an equal amount of space, image flattening would be necessary. The class activation maps technique will be utilized by the author to determine which part of the image was used to categorize the specific class of emotion using the Convolutional neural network. After the model has been fully trained for emotion identification, an external picture will be input into the educated algorithm, allowing the real-time image to be classed using the same technique. The final model is expressed using graphs and different visualizations, making it easy for both the author and the reader to understand just by looking at the visuals of the models and modeled photos.
In the very first step, the author has done a dictionary mapping for all seven emotions. Thereafter, set the size for each image as 20 x 30 inches and then plot a simple raster image for each sub-plot using the enumerate function for each axis.

3.2 Pre-processing

CNN performance may be influenced by a variety of circumstances, including crowded backgrounds, lighting, and postural variation. The use of preprocessing filters may increase the accuracy of identifying facial expressions. Sharpening photos, for example, can improve the edges of vital features like the lips and eyes. These contours are critical in predicting face emotions. When the colors are the same, histogram equalization aids in distinguishing the foreground from the background. Three pre-processing strategies commonly employed in facial expression recognition models are used in our research. In data preprocessing author first convert the single-channel image to a 3-channel image i.e., simply convert the dataset image from grayscale to RGB image. For training our model the author keeps 80% of the total images and the rest 20% for testing our trained model. After that, reshape the train set, validation set, and test set to a 48x48x1 dimensional array.

3.3 Data Augmentation

The first method is data augmentation, which was used as a preprocessing step in all of our models. For data augmentation, the author has utilized Keras' "ImageDataGenerator." As

demonstrated in Figure 2, it creates 32 augmented pictures from a single image by rotating, flipping, and using other specified processes.

Figure 2: Extended each image to five images using transformations.

4 Design Specification

4.1 Training Architecture

After pre-processing, the proposed model was trained using 80% of the pre-processed images from the Facial Expression Recognition Challenge Dataset. For each subject, there are seven standard types of emotions in this database. For feature extraction and classification purpose, the author has used the Deep Convolutional Neural Networks throughout the training phase. It employs a supervised learning method across a large number of photos. Figure 3 depicts the proposed convolutional neural network (CNN).

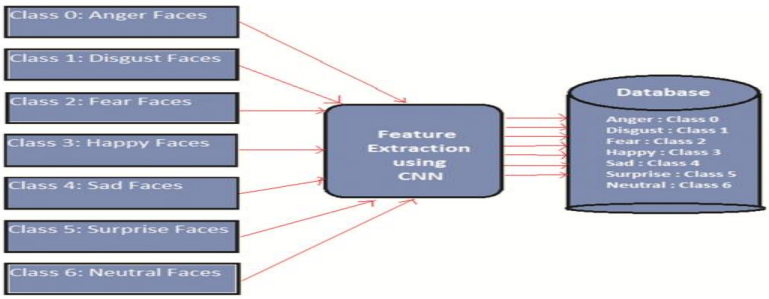

Figure 3: Training Architecture

4.2 Testing Architecture

For testing the model, the author has chosen 20% of the photos from the Facial Expression Recognition Challenge Dataset after pre-processing. In the testing stage, feature extraction is performed using the convolutional neural network, as in the training phase. However, emotion categorization is determined after comparing extracted features with taught features.

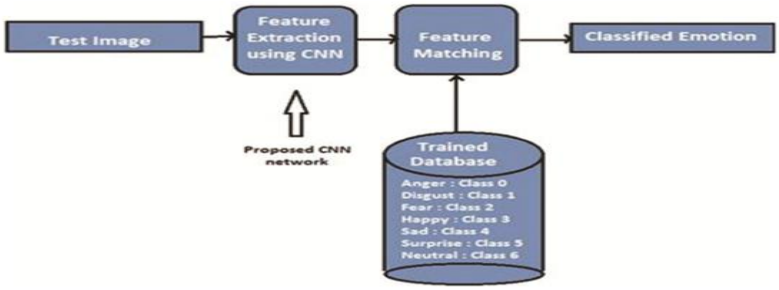

Figure 4: Testing Architecture

4.3 Convolutional model

The most impactful advancements in the field of computer vision have been convolutional neural networks. It is physiologically inspired by the visual cortex and mimics how the visuals are been processed by the human brain. Keras library has dominantly presented this entire study with the help of the TensorFlow environment which allows the python packages and this made it possible to conduct experiments more quickly and easily. ConvNet designs explicitly assume the default input as an image and supervise particular architectural methods which subsequently analyze the image's characteristics. ConvNet's layers are summarized as follows:

4.3.1 Input Layer

In the Input layer, there is the pixel value of the image in the form (H x W x C), where H represents the width of the image, W points to the width of the image and C denotes the number of channels of colors. In this study, it is formulated as (48 x 48 x 1), where the color channel is 1 which represents grayscale photos. Because the dimensions are fixed, pre-processing is required for the pixels so that they can be processed in the Input layer.

4.3.2 Convolutional layer

The Convolutional layer offers the product of the pixel weight of the image with the small space to which the convolutions are associated in the Input layer. The most exclusive feature value is the number of filters utilized in the convolutional layer. It generates a random weight to the filter which acts like a hyperparameter to this layer. This filter can also be referred to as kernel which convolutes the multiplication of the random weight of the element with the pixel value of the input image. This dot product gives us the feature map which a distinct feature ID for each element and helps in finding the orientation of the image with the associated edges to it. This feature ID further helps in improving the pixel value of the image which eventually results in a better image recognition system. The dot product generated for feature ID is represented as (w * h * f), where f represents the number of filters employed. Due to the increase in convolutional layers, the computational time is also increased for dimension reduction. To overcome this challenge, the pooling layer is brought into the picture to help consecutive dimensionality reduction along the height and the width of the image. MaxPooling layer helps in reducing the dimension mapping by a window size and holds the maximum pixel value of the original feature mapping window.

4.3.3 Fully connected layer

The output of the pooling layer is linked as an input to this fully connected layer which is also known as the dense layer. These outputs are mostly applied at the last stage of the Convolutional Neural Network to know the exact number of outputs to the model. This helps in correcting the features associated with the layers having random training weights. This layer benefits in improving the complex features of the image which results in precised feature mapping of the entire image. But during this time it can be likely to overfit the model and to prevail over this task, a dropout layer is included at the end of this layer which picks apart (typically less than 50%) of nodes during training and sets their weights to zero.

4.3.4 Output layer

The output layer is connected with the outputs of the fully connected layer giving out the probability and the classification of the image. As some emotions in humans are often a combination of emotions, the likelihood of each emotion is determined. This is accomplished by utilizing the network's SoftMax layer.

4.4 VGG Network

In 2014, Karen Simonyan and Andrew Zisserman from the University of Oxford created the VGG network which then became an integral part of the Convolutional Neural Network. In this neural network, there's an input having 3 channeled RGB pixel label which measures 224 x 224 pixels. This design employs a Conv3 layer which has a filter size of 3 x 3 and a Conv1 layer having a filter size of 1 x 1(Aza et al. 2020). The size of the convolutional layer utilized varies, as indicated in Figure 5, from 64 x 64 to 128 x 128, 256 x 256, and 512 x 512.

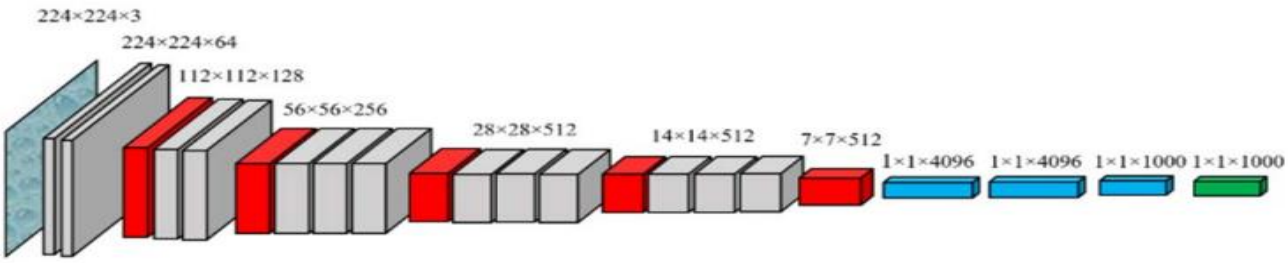

Figure 5: VGG Architecture

5 Implementation

For face detection, the author utilized the OpenCV library. The Viola-Jones Haar-like detector (P. Viola et al. 2004) serves as the foundation for the face detection method. Because it uses integral pictures, which decrease duplicate operations, this approach is extremely fast.

a. Creating a dataset for emotion classification using a convolution neural network layer.
b. Making a database using an open CV library.
c. Recognition and classification training procedure.
d. All files are being built for Anaconda 3.4 and Python 3.7 software platforms.

Below you can see the architecture of both the CNN models used in the implementation of the model. On the right, there's a Convolutional model with 3 convolutional layers and on the left, you can see the complete and broad architecture of the Convolutional model with 5 convolutional layers.

CONVOLUTIONAL 5

Conv2d_1_input: InputLayer
Conv2d_1: Conv2D
batch_normalization_1: BatchNormalization
conv 1: Conv2D
batch_normalization_2: BatchNormalization
max_pooling2d_1: MaxPooling2D
dropout_1: Dropout
conv2: Conv2D
batch_normalization_3: BatchNormalization
conv3: Conv2D
batch_normalization_4: BatchNormalization
conv4: Conv2D
batch_normalization_5: BatchNormalization
max_pooling2d_2: MaxPooling2D
dropout_2: Dropout
conv5: Conv2D
batch_normalization_6: BatchNormalization
conv6: Conv2D
batch_normalization_7: BatchNormalization
conv7: Conv2D
batch_normalization_8: BatchNormalization
conv8: Conv2D
batch_normalization_9: BatchNormalization
max_pooling2d_3: MaxPooling2D
dropout 3: Dropout
conv9: Conv2D
batch_normalization_10: BatchNormalization
conv10: Conv2D
batch_normalization_11: BatchNormalization
conv11: Conv2D
batch_normalization_12: BatchNormalization
conv12: Conv2D
batch_normalization_13: BatchNormalization
max_pooling2d_4: MaxPooling2D
dropout 4: Dropout
conv13: Conv2D
batch_normalization_14: BatchNormalization
conv14: Conv2D
batch_normalization_15: BatchNormalization
conv16: Conv2D
batch_normalization_16: BatchNormalization
conv17: Conv2D
batch_normalization_17: BatchNormalization
max_pooling2d_5: MaxPooling2D
dropout_5: Dropout
flatten_1: Flatten
dense_1: Dense

CONVOLUTIONAL 3

Conv2d_1_input: InputLayer
Conv2d_2: Conv2D
Conv2d_3: Conv2D
max_pooling2d_6: MaxPooling2D
dropout_6: Dropout
Conv2d_4: Conv2D
max_pooling2d_7: MaxPooling2D
Conv2d_5: Conv2D
max_pooling2d_8: MaxPooling2D
Conv2d_6: Conv2D
max_pooling2d_9: MaxPooling2D
dropout_7: Dropout
flatten_2: Flatten
dense_2: Dense
dropout_8: Dropout
dense_3: Dense

6 Evaluation

The author creates and tests the facial expression recognition technique using the Jupyter Notebook. Windows 10 is the operating system in use. The cloud utilized includes Intel Core i7 (10th Gen) characteristics with 8 GB of Random-Access Memory (RAM), a 1.5 GHz processor, and an Intel Iris Plus 2 GB Graphics Processing Unit (GPU).
At the training stage, the FER Challenge dataset is divided in the ratio 80:20 percent for training and testing respectively.
As a result, there are 22967 photos for training and 5741 for testing. On the FER Challenge dataset, the recognition objective consists of seven emotion classes: angry, happy, surprise, fear, sad, disgust, and neutral. Then, training architectures using 5 convolutional layers and 3 convolutional layers, as well as VGG16, using initially trained model as parameter standards. Adam, RMSprop, and Adam are the optimization algorithms employed, and the learning rates are 0.0001, 0.0001, and 1e-10 respectively for the three models. Figures 6, 7, and 8 demonstrate the accuracy obtained in the training phase for all of the three models having Adam, RMSprop, and Adam functions for optimization of the model, respectively. The accuracy graph is rather steady when Adam is used for the 5 convolutional layered models, even though it employs a small epoch.

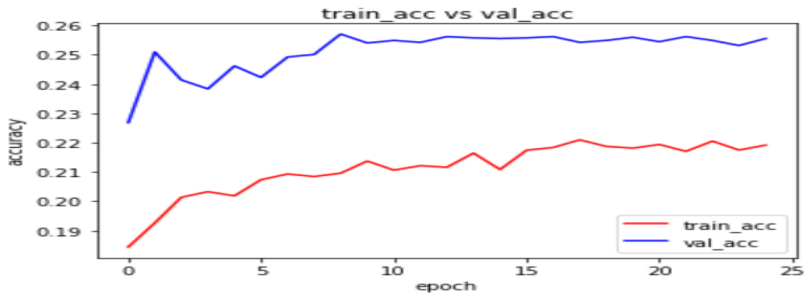

Figure 6: Training accuracy for Convolutional 5 model

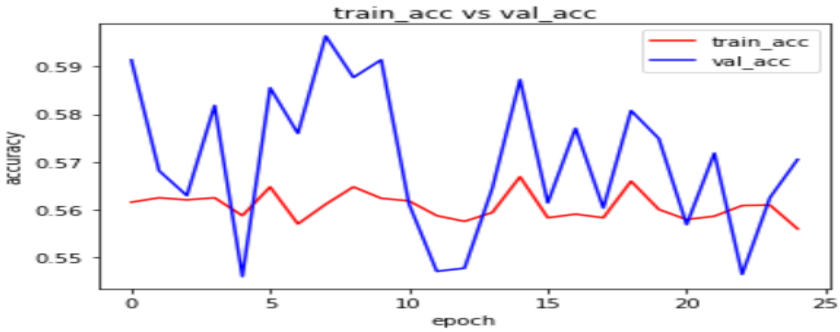

Figure 7: Training accuracy of Convolutional 3 model

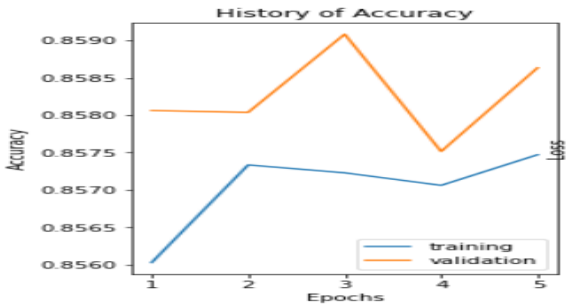

Figure 8: Training accuracy of VGG16 model

The proposed models built during the training phases are employed for recognizing the facial expression from each image depending upon the different emotion classes. Each model is compared based on the test loss, test accuracy, validation accuracy, and validation loss. Table I shows the evaluation findings. The experiment reveals that the VGG16 architecture trained with a learning rate of 1e-10 and Adam optimizer delivers the greatest results, achieving 85.75 percent accuracy, 49.42 percent precision, 2.41 percent recall, and 4.26 percent F1-score.

Table 1: Results Evaluation

Model	Accuracy (%)	Loss
Convolutional 5	25.6	1.78
Convolutional 3	57.2	1.18
VGG16	85.7	1.73

Figures 9 and 10 demonstrate the confusion matrix for CNN 5 and CNN 3 architectures with Adam and RMSprop optimizers and a learning rate of 0.001.

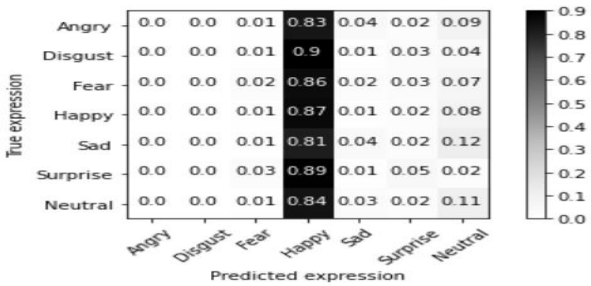

Figure 9: Confusion matrix for Convolutional 5 model

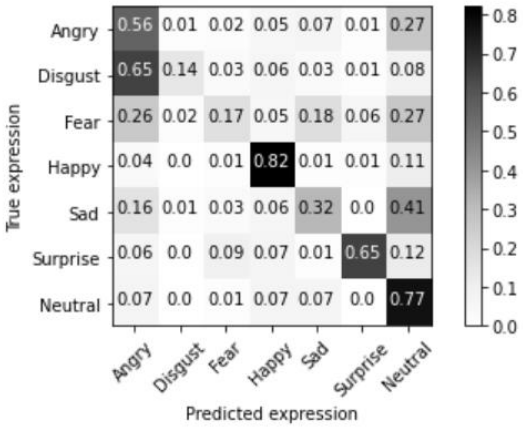

Figure 10: Confusion matrix for Convolutional 3 model

Fear, sadness, happiness, and surprise are all expression classes that have recognition problems. One image with true class anger, for example, is incorrectly classified as sad, and two photos with the neutral class.

7 Ethics

The dataset utilized for the proposed study is open-source data from the Kaggle website, which does not raise any ethical concerns about data privacy or other elements of data usability. As a result, there is no requirement for an ethical statement form in this study activity.

8 Conclusion and Discussion

In this research, the author has presented 3 models for evaluating the result. For Convolutional 5 and Convolutional 3 models, results might not seem promising but in the VGG16 the final result provides the best accuracy. There is still a chance of improvement from the above 3 models. In some classes of emotions, there is confusion between the emotions which is a result of data imbalance. Most of the computer vision technologies like facial expression recognition and characteristic forecasting exhibit class imbalance. Deep Convolutional techniques used today usually employ element re-sampling or cost-efficient learning strategies. The author proved its use via extensive tests and showed that the suggested Cluster-based Large Margin Local Embedding (CLMLE) performs astonishingly well when linked with the basic k-nearest cluster classifier. CLMLE preserves inter-cluster angular margins both inside and across classes, resulting in significantly more balanced class borders at the local level. Our feature learning produces new state-of-the-art performance on current face recognition and attributes benchmarks in a short period.
To improve feature discrimination, LMCL generalizes SoftMax by requiring a large angular gap between classes. If this does not work, the author shall resort to resampling which is used to deal with data imbalance. It overcomes this imbalanced data, it either executes under-sampling in which the imaged are deleted from the mainstream class, or else it performs over-sampling in which the number of images in the submissive class is added at each instance.

9 Acknowledgment

First and foremost, I'd want to thank our module lecturer, Dr. Catherine Mulwa, and project supervisor Mr. Jorge Basilio, for their tireless efforts in making the module run smoothly during the semester. It would not have been feasible without your help. Repeated efforts were made to describe every detail of the project proposal report. It was a pleasant and educational experience working with the professors and my classmates to overcome any obstacles I encountered along the way.

10 References

1. A. Fathallah, L. Abdi and A. Douik, "Facial Expression Recognition via Deep Learning," *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, 2017, pp. 745-750, doi: 10.1109/AICCSA.2017.124.
2. Aza, M.F.U., Suciati, N. and Hidayati, S.C., 2020, October. Performance Study of Facial Expression Recognition Using Convolutional Neural Network. In *2020 6th International Conference on Science in Information Technology (ICSITech)* (pp. 121-126). IEEE.

3. Baron-Cohen, S., & Tead, T. H. E. (2003) Mind reading: The interactive guide to emotion. London: Jessica Kingsley Publishers.
4. Castellano G., Kessous L., Caridakis G. (2008) Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech. In: Peter C., Beale R. (eds) Affect and Emotion in Human-Computer Interaction. Lecture Notes in Computer Science, vol 4868. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3- 540-85099-1_8
5. Castellano G., Villalba S.D., Camurri A. (2007) Recognising Human Emotions from Body Movement and Gesture Dynamics. In: Paiva A.C.R., Prada R., Picard R.W. (eds) Affective Computing and Intelligent Interaction. ACII 2007. Lecture Notes in Computer Science, vol 4738. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74889-2_7
6. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. arXiv preprint arXiv:1409.4842, 2014.]
7. D. Glowinski, A. Camurri, G. Volpe, N. Dael and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008, pp. 1-6, doi: 10.1109/CVPRW.2008.4563173.
8. E. Bal, E. Harden, D. Lamb, A. Van Hecke, J. Denver, and S. Porges. Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state. Journal of Autism and Developmental Disorders, 40(3):358– 370, 2010.
9. F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera and G. Anbarjafari, "Survey on Emotional Body Gesture Recognition," in IEEE Transactions on Affective Computing, vol. 12, no. 2, pp. 505-523, 1 April-June 2021, DOI: 10.1109/TAFFC.2018.2874986
10. Gervasi, Osvaldo et al. ‘Automating Facial Emotion Recognition. 1 Jan. 2019: 17 – 27.
11. LIANG, Yanqiu. Intelligent Emotion Evaluation Method of Classroom Teaching Based on Expression Recognition. International Journal of Emerging Technologies in Learning (iJET), [S.l.], v. 14, n. 04, p. pp. 127-141, Feb. 2019. ISSN 1863-0383. Available at: Date accessed: 14 Aug. 2021. doi:http://dx.doi.org/10.3991/ijet.v14i04.10130.
12. Ly, Son Thai; Lee, Guee-Sang; Kim, Soo-Hyung; Yang, Hyung-Jeong (2018). [ACM Press the 2018 International Conference - Ha Noi, Viet Nam (2018.09.28- 2018.09.30)] Proceedings of the 2018 International Conference on Machine Learning and Machine Intelligence - MLMI2018 - Emotion Recognition via Body Gesture. , (), 27–31. doi:10.1145/3278312.3278313
13. Mehta, D.; Siddiqui, M.F.H.; Javaid, A.Y. Facial Emotion Recognition: A Survey and Real-World User Experiences in Mixed Reality. Sensors 2018, 18, 416. https://doi.org/10.3390/s18020416
14. Metcalfe, Dale; McKenzie, Karen; McCarty, Kristofor; Pollet, Thomas V. (2019). Emotion recognition from body movement and gesture in children with Autism Spectrum Disorder is improved by situational cues. Research in Developmental Disabilities, 86(), 1–10. doi:10.1016/j.ridd.2018.12.008
15. Mishra S., Prasada G.R.B., Kumar R.K., Sanyal G. (2017) Emotion Recognition Through Facial Gestures - A Deep Learning Approach. In: Ghosh A., Pal R., Prasath R. (eds) Mining Intelligence and Knowledge Exploration. MIKE 2017. Lecture Notes in Computer Science, vol 10682. Springer, Cham. https://doi.org/10.1007/978-3-319-71928-3_2

16. M. Lin, Q. Chen, and S. Yan. Network in network. arXiv preprint arXiv:1312.4400, 2013.
17. Mollahosseini, A., 2018. *Developing an Affect-Aware Rear-Projected Robotic Agent* (Doctoral dissertation, University of Denver).
18. Mollahosseini, A., Chan, D. and Mahoor, M.H., 2016, March. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)* (pp. 1-10). IEEE.
19. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. arXiv preprint arXiv:1409.0575, 2014
20. Piana, Stefano; Staglianò, Alessandra; Odone, Francesca; Camurri, Antonio (2016). Adaptive Body Gesture Representation for Automatic Emotion Recognition. ACM Transactions on Interactive Intelligent Systems, 6(1), 1– 31. doi:10.1145/2818740
21. P. Viola and M. J. Jones. Robust real-time face detection. International journal of computer vision, 57(2):137–154, 2004.
22. Santhoshkumar, R.; Geetha, M. Kalaiselvi (2019). Deep Learning Approach for Emotion Recognition from Human Body Movements with Feedforward Deep Convolution Neural Networks. Procedia Computer Science, 152(), 158– 165. doi:10.1016/j.procs.2019.05.038
23. Shi, J.; Liu, C.; Ishi, C.T.; Ishiguro, H. Skeleton-Based Emotion Recognition Based on Two-Stream Self-Attention Enhanced Spatial-Temporal Graph Convolutional Network. Sensors 2021, 21, 205. https://doi.org/10.3390/s21010205
24. Shunyi Wang. Online Learning Behavior Analysis Based on Image Emotion Recognition. Traitement du Signal. 2021;38(3):865-873. doi:10.18280/ts.380333
25. S. Saha, S. Datta, A. Konar and R. Janarthanan, "A study on emotion recognition from body gestures using Kinect sensor," 2014 International Conference on Communication and Signal Processing, 2014, pp. 056-060, doi: 10.1109/ICCSP.2014.6949798
26. Suchitra, Suja P. and S. Tripathi, "Real-time emotion recognition from facial images using Raspberry Pi II," 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), 2016, pp. 666-670, doi: 10.1109/SPIN.2016.7566780.
27. S. Arora, A. Bhaskara, R. Ge, and T. Ma. Provable bounds for learning some deep representations. arXiv preprint arXiv:1310.6343, 2013.
28. S. Jerritta, M. Murugappan, R. Nagarajan and K. Wan, "Physiological signals based human emotion Recognition: a review," 2011 IEEE 7th International Colloquium on Signal Processing and its Applications, 2011, pp. 410-415, doi: 10.1109/CSPA.2011.5759912.
29. U. Agrawal, S. Giripunje and P. Bajaj, "Emotion and Gesture Recognition with Soft Computing Tool for Drivers Assistance System in Human Centered Transportation," 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 4612-4616, doi: 10.1109/SMC.2013.785
30. Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. CoRR, abs/1502.00873, 2015.
31. Yang, B., Li, Z., Sun, Y. and Cao, E., 2020. EM-FEE: an efficient multitask scheme for facial expression estimation. *Interacting with Computers*, *32*(2), pp.142-152.