MENTAL HEALTH RECOGNITION USING ONE-SHOT LEARNING AND FACIAL EXPRESSIONS

Mahesh Kumar N¹, Aditi Das², G. Tharun³, P. Sujith⁴
¹Assistant Professor/ECE, ²,³,⁴B. Tech in ECE, ¹,²,³,⁴ SRM Institute of Science and Technology

ABSTRACT

The visual representation of a person's affective state, cognitive function, purpose, personality, and psychology can be obtained from facial expressions. The idea of our project is to create a platform with a camera, to determine whether an individual is suffering from deterioration of mental health, particularly in an institutionalized environment. In general, disturbed mental condition is determined by using a combination of various factors such as facial expressions, body posture, speech, tone, gestures, and so on. This project lays focus on Facial Expression Recognition (FER), achieved with the help of Convolution Neural Network (CNN), One-Shot learning and TensorFlow. The dataset used here is the Kaggle dataset FER2013, which comprises of seven facial expressions labelled as happy, sad, surprise, fear, anger, disgust and neutral. The results of this classification are stored and compared to determine whether an individual displays negative emotion for a significant number of times. After detection and classification, a prompt is sent to the designated authorities at the institution through a web application. On the basis of that, if an individual is found to be suffering from a disturbed state of mind, he/she is recommended for therapy or medical attention.

Index Terms—Mental Health, Facial Expression, Depression Detection, Emotion Recognition

I. INTRODUCTION

Mental health problems, although always prevalent, have seen a significant spike in the recent decades. Suicide kills a devastatingly large number of people each year, making it the second leading cause of death among teenagers. People with such conditions lead dissatisfied lives, consistently feeling sad or hopeless beyond control. Such people then start retrieving from their social lives and stop indulging in their usual activities. The overall quality of their life is reduced significantly. Even students at many instances find it very stressful to handle the intense academic pressure, resulting in stress-related problems from a very young age. Several mental conditions can also happen due to an abrupt lifestyle change, such as the outbreak of the 2019 COVID pandemic, which resulted in people being confined within their houses for months, left alone with their harmful thoughts, without any contact with their loved ones. Parallely, due to the rise of social networking sites, teenagers are constantly in a state of self-doubt – questioning whether they are socially acceptable by their peers. Social media sites create a sense of false belief that others are leading more successful and happier lives. Furthermore, spending more time online may make one a victim of cyber-bullying, which can in turn cause mental health conditions. In addition to this, certain online games like "The Blue Whale Challenge" have been known to target people who are already in an emotionally vulnerable state of mind.

In most situations, individuals who are depressed are totally ignorant of their disturbed mental condition. They are unable to identify the cause of the constant unhappiness within them, allowing their minds to wander towards life-risking tendencies. Sometimes, people know that they are suffering from mental conditions, but they are hesitant in seeking any kind of help, mainly due to ill-conceived notions of humiliation associated with any kind of mental problem. In countries like India, it is predominantly a taboo. Thus, it is better to identify the signs at initial stages. Mental conditions, if identified early, can be eased with a simple session with a counsellor. This may be of immense help, and can significantly change the negative state of mind of the person into a positive one. They can be helped with ways to deal with mental stress and can be made to feel like they are not alone in that dark space. Numerous studies have detected that people with different mental illnesses exhibit different physiological and behavioral indicators. Thus, the neural activity behind different emotions on faces...
can help distinguish between such conditions. Machine learning models use this aspect to diagnose various complex psychological symptoms of mental disorders[6]. Facial expressions of individuals are the easiest form of biometry to obtain. They can be analyzed and the emotion computed to detect people with potential mood disorders. This has the ability to aid in the development of a practical method for evaluating multi-dimensional prediction models [8]. This can be useful in various situations for person-monitoring at organizations.

II. RELATED WORK

Decoding people's feelings has been a source of debate for thousands of years, dating all the way back to Aristotle in the 4th century BC. The history of this domain is not surprising. Emotions are part of all of us, but the vast majority of conclusions taken from it are focused on hypotheses that lack empirical support. This lack of scientific explanation has sparked a slew of controversies and competing hypotheses all the way up to the present day.

Using One-Shot learning for Recognition

Humans display an innate ability to recognize patterns and understand variations. Similarly, machine learning algorithms have achieved superior results in a number of applications, including web search, spam detection, and speech and image recognition. But these algorithms perform well only when they are trained with extensive supervised information, and generally fail when trained with small datasets. Several experiments have been carried out in order to popularize these unfamiliar categories so that intensive retraining, which could be costly or impractical due to circumstances, can be avoided, without causing any resistance in the accuracy of the output. But this algorithm is fairly recent as compared to other CNNs. A study has been conducted, which outlined new findings comparing network output to a state-of-the-art classifier designed for the Omniglot data set [a]. The networks outperformed all other models by a significant margin, coming close to the best results previously achieved. It was claimed that these networks' strong performance on the task suggested that human-level accuracy was possible, and that the method should be applied to one-shot learning tasks in other domains, especially image classification.

Facial Expression Recognition

A lot of valuable information can be gotten from facial expressions, which provides a useful method of determining a person's level of consciousness and mental activity[1]. Due to this, the identification of facial expressions, which has significant theoretical, functional, and life application importance, has become a hot topic of study. Depending on people’s general way of thinking being reflected on their face, facial expression recognition employs a computer to extract features, analyze facial expression data, and comprehend human emotions such as happiness, surprise, rage, fear, disgust, and sadness, among others.[9]. Many articles discuss recent developments and applications in facial expression recognition, ranging from face identification to feature extraction to classification to ethnic expression recognition [9].

Recent researches suggest that experiments investigating emotional processes should be focused on the evaluation of images containing effective and useful information, such as facial expressions. Databases containing emotion-inducing stimuli may help researchers in the future, where the stimuli's attributes that can affect the outcome can be regulated. One such research presented data from Portugal for the detection of seven emotional facial expressions along with a neutral face, on the Radboud Faces Database (RaFD). The gender of subjects and the sex of models had an effect on emotion recognition. 312 images of white people in their youth with emotional and neutral expressions and a frontal gaze were shown to participants (N = 1249). The degree of alignment between the demonstrated and desired expressions ranged from 69 percent (for anger) to 97 percent (for happiness). Only for anger and scorn were women's recognition levels substantially higher than men's. [10]. Overall, the findings indicated that the RaFD offered adequate stimuli for experiments investigating the identification of emotions and facial expressions, and their recognition.

Another study found that facial emotion recognition is regulated by specific neural substrates and is an essential feature of interpersonal communication [11]. Furthermore, the most recent literature on emotion recognition in psychological and neurologic disorders is examined and analyzed in order to compare and contrast various disorders[10]. Yet another research suggests using Ensemble Convolutional Neural Networks to detect depression in speech automatically. This paper describes an automated method for detecting depression by analyzing a person's speech. This deals with lot features like energy, pitch and voice quality of the speech.
the extracted data is merged and then at the end it detects the type of emotion with the help of the following data.

[12].

III. PROPOSED SYSTEM

Convolutional Neural Networks (CNN) can be used for a variety of tasks, including natural language processing, speech recognition, object detection, and so on, but they are most commonly used to solve image processing problems. One of the key reasons why deep learning and artificial intelligence are so important today is because they are so useful in these fast-growing fields.

Convolution Neural Network

CNN is more effective in processing and classifying images as compared to any other neural networks. A CNN is made up of multiple layers[12][13], and contrary to other neural networks, they have a feature extractor, which is comprised of a convolution layer and a down-sampling layer. CNN is generally made up of three basic layers: input, hidden, and output. The first layer (input layer) is the unprocessed original image that is fed into the model. The output layer is the final product after classifying the features using the hidden layer and neuron layers with a multiple nonlinear structures, including a convolution layer and a sub-sampling layer. In hidden layers, CNNs extract and identify features[13]. They also contain the Softmax function, which is an activation feature that ensures non-linearity as data passes across the network layers. Without it, the image's dimensionality will be lost as data was fed into each sheet. We can perform classification on the dataset, thanks to the completely connected layer.[14][15].

Deep Learning

Deep learning enables the machine to mimic human behavior. It achieves this through algorithms trained with datasets. For example, a human can easily identify a number written on a paper. But when a computer has to recognize it, that's where deep learning comes into play. The neural network is equipped to recognize the handwritten numbers as an image of NxN pixels, for example: (48x48), (64x64). Neurons are the building blocks of the neural network, where information is processed. The input layer of the neural network is formed by feeding each of the total pixels to the neuron in the first layer. The output layer, with each neuron representing a digit, is on the other end, with information transferred from one layer to the next through connecting channels. A weighted channel is one that has a value attached to each of these. Bias is a number that is correlated with each and every neuron. This bias is then multiplied by the weighted number of inputs hitting the neuron, which is then added to a function known as the activation function. The effect of the activation function decides whether or not a neuron sends information to the next layer. To create a well-trained network, the weights and biases are continuously balanced. In terms of resolving problems, deep learning has made substantial progress, eliminating factors that had been hampering the artificial intelligence community's best efforts for many years. In image recognition, deep learning-based methods have outperformed traditional machine learning approaches in terms of accuracy and processing speed.

Residual Neural Network (ResNet)

ResNet is a form of neural network that can learn to recognize images and develop over time. It utilizes skip connections to hop over different layers, and is usually implemented with double and triple layer skips with varying nonlinearities. There are two main reasons to skip connections: to avoid vanishing gradients and to mitigate the degradation problem. In the simplest case, the weights for the adjacent layer are adapted to minimize the distortion caused by the upstream layer. Skipping allows a network to simplify its training by reducing the number of layers that it uses. It also helps minimize the impact of vanishing gradients by keeping the network closer to the manifold. Furthermore, they allow models to learn identity functions, ensuring that the upper layer outperforms the lower layer. In order to boost accuracy and performance, of Deep Neural Networks, more layers are stacked. This allows the layers to learn more complex features and solve more complex problems.

One-Shot Learning

One-Shot Learning is a classification or object categorization model in which one or a few examples are used to classify a large number of new ones. It is a form of computer vision that can compare two images that it has never seen before and tell if they are of the same object[16]. This refers to a variety of work in the field of facial recognition, where people must be correctly categorized based on one or only few template images, despite having different facial expressions, lighting conditions, clothing, hairstyles and so on. The challenge of one-shot learning is addressed in advanced face recognition systems by acquiring a dense low-dimensional feature
representation known as a face embedding, which can be easily calculated and measured for verification and identification purposes. The Siamese network is a network that has gained popularity as a result of its use of one-shot learning. A Siamese network combines the outputs of two parallel neural networks into a single network, where each requires a unique set of inputs. Siamese networks are a method for dealing with one-shot learning that compares a trained feature vector for the identified candidate. The verification model learns to classify input pairs based on their likelihood of belonging to the same or different groups. This model can then be used to pairwise evaluate new images against the test image, one for each novel class. For the one-shot mission, the pairing with the highest score according to the verification network is given the highest likelihood.

IV. METHODOLOGY

Face recognition has regained popularity following the rapid growth of artificial intelligence due to its nonintrusive existence. When compared to other forms of biometric techniques, it is the most common form of human identification [17]. In an unregulated setting, face recognition can also be easily checked without the individual's knowledge. Face recognition has been studied and discussed in a number of research papers throughout history [18]–[23]. Pose variance, facial disguises, lighting of the scene, the sophistication of the picture context, and shifts in facial expression have all been problems for traditional approaches based on shallow learning [24]–[31]. To extract sample features, shallow learning methods depend on artificial experience, and only use a few basic features of images. Therefore, methods based on deep learning can retrieve much more complex facial features.

This paper proposes an updated Convolutional Neural Network (CNN) architecture by using normalization operations to separate distinguishing facial features, and a Softmax classifier to identify the emotions in those faces. CNNs are deep neural networks designed from biologically driven structures which are found in the human brain visual cortex. In general, a CNN can be thought of as an artificial neural network with some specialization in the ability to identify or pick out patterns and make sense of them. They have hidden layers known as convolutional layers, and these layers are what allow them to work. A CNN may have non-convolutional neural networks as well. A convolutional layer, like every other layer, receives input and then transforms it in a specific way to produce an output. This is then converted into the next layer's input. This happens multiple times to form a convolutional operation. It is possible to determine the number of filters that each convolutional layer should have. These filters detect patterns. This project captures image frames from cameras installed in various locations of common gathering at institutions, such as cafeterias, water coolers, etc. The images are then sent to the deep learning algorithm comprising of the following layers:

Convolution 2D layer: Sliding convolutional filters make up a 2-D convolutional layer. It moves the filters along the input vertical and horizontally. This analyses the face and determines which parameters to consider for expression classification.

Batch Normalization Layer: The analysis of data through several observations is made easier with a batch normalization layer. It can also reduce the sensitivity of network initialization by keeping the data in its original state.

Activation layer: The activation function is a key parameter of the CNN model. It decides which information should be disseminated in the network and which should be retained.

Max-pooling 2D layer: By dividing the input into different rectangular pooling regions, a Max pooling layer performs down-sampling. This is accomplished by the use of convolutional layers.

Dropout layer: When all the features of the model are connected to the same network, this can cause overfitting of the training dataset. To overcome this issue, a dropout layer is used. This layer consists of dropping some neurons from the neural network during the training process. This ensures that over-training of the model is not done and provides a more accurate result.

The One-Shot Learning model is trained using few images of every individual of the organization, so that the model can continue to determine the mental health for them. The CNN model is trained using the Kaggle dataset [FER2013]. Softmax function is used at the output end to evaluate the output of a neural network and normalize the probability distribution of each class. This function returns a vector containing the probability distribution of a set of possible outcomes. The neural network will display the output class with highest probability value.
The integration of one-shot learning with CNN is to ensure that the images of the subject are captured when they are not conscious of being evaluated. Obtaining their images at random intervals allows for more genuine emotions.

**Figure 1. Architecture**

**Performance of similar models –**

Happiness is the most easily detected emotion, according to research, even at low intensities, with accuracy reaching up to 100 percent [(32]. The rating of the remaining universal emotions has differed across research, which may be due to various test formats that use different types and intensities of emotions, exposure length, number of stimuli, and other factors. It was observed that people who were in a decent state of mind displayed positive expressions for the maximum span of time, whereas those who were battling with some internal struggle failed to do so. Neutral expressions are difficult to comprehend, considering that they can be implied in any way. But when observed for longer spans of time, the other expressions show up, making the process of classification into mentally stable or not more determinable.

**Figure 2. Flow Diagram**

**V. EXPERIMENTAL RESULTS**

**One-Shot Learning Output**

On running the images taken by the camera through the one-shot learning neural network, the individual identification is achieved. The frames captured in each shot of the camera is processed by the neural network, which compares it with the pre-existing images in the database. It then creates separate folders with the names or IDs for each face that it detects. After that, every time an individual’s image is captured, their image is stored into that folder. This acts as the computer vision for our model.
The above output is of individuals whose photographs were used to train the model with. Thus, the system is able to identify and name them. Simultaneously, it creates a folder for every individual where it stores all their images at one place.

The above person is a completely new individual whose image was not used to train the system. That is why the system is unable to identify the person.

**Convolution Neural Network Output**

After person recognition is done using one-shot learning, the image is submitted to the CNN model, which recognizes and classifies the emotion into one of the seven facial expressions - happy, sad, surprise, fear, anger, disgust and neutral. The emotions are then classified into positive (happy, surprise and neutral) and negative (sad, fear, anger and disgust). Several frames are captured from the camera and processed in the same way to obtain multiple images of each person as they pass by the camera. A threshold of 0.8 is then set to determine the mental health condition of the person such that, if 80 percent of the images captured of an individual display negative emotions, that person is declared in need of medical attention. If it is found that an individual displays negative emotions less than that, they are considered to be in a normal state of mind. Finally, a prompt can be sent to the authorities of the institution, informing them of the consistent mental state of individuals belonging to their database. The authorities can then choose to take the appropriate action on the basis of this information. Counselling sessions can be arranged, yoga or meditation time can be allotted for those in need of them, and external therapy can also be recommended if nothing else helps.
VI. CONCLUSION AND FUTURE WORK

It is imperative to understand that mental health problems are not black and white, but the gray areas in between. It is something that every individual faces in their life, with different intensities. An organization performs best when its personnel are physically and mentally fit. Although we cannot diagnose a person using just a few photos, this project aims to offer a high level recognition of those people who need help, and acts as a medium to encourage them to do so, ascertaining a better quality of life.

This model can further be turned into a computer software, which clicks random images through the webcam and processes them to determine the mental well-being of individuals. This can help to accommodate the work from home culture that has become a necessity in the recent days. In the future, upon merging with physiological signals, this project has the capability of providing results for medical use with much greater accuracy.

REFERENCES

2. Michael Hurst. Emotion Recognition – Theory or Practicality. 18th International Conference
3. on Automation & Computing, Loughborough University, UK [2012]