A Multimodal Approach for Emotion Detection through Facial Expressions, Recommending Movies and Songs Using Deep Learning Model

MSc Research Project
Data Analytics

Rajshri Pawar
Student ID: 22126571

School of Computing
National College of Ireland

Supervisor: Taimur Hafeez

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Rajshri Pawar
Student ID:	22126571
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Taimur Hafeez
Submission Due Date:	14/12/2023
Project Title:	A Multimodal Approach for Emotion Detection through Facial Expressions, Recommending Movies and Songs Using Deep Learning Model
Word Count:	2567
Page Count:	22

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:	
Date:	31st January 2024

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):	

A Multimodal Approach for Emotion Detection through Facial Expressions, Recommending Movies and Songs Using Deep Learning Model

Rajshri Pawar
22126571

Abstract

The term ”personality” describes a variety of traits, including emotions playing a big role, ideas, and social behaviours, that collectively give an individual their unique character. This paper explores a novel approach to Emotion detection and suggesting movies and songs according to the emotion that goes beyond traditional techniques. Through a combination of psychological, creative, and machine learning insights, our novel methodology goes into the unexplored realm of deep learning methodologies. Rather than relying on conventional techniques, we utilize a combination of transformer-based models which allows us to explore the subtle visual details that influence Emotion. Our method differs from others since we are including facial expression images and recommending movies and songs. We recognize that Emotions is a multidimensional feature of human expression, going beyond the limitations of text-based data. We use photos as a crucial data source in Emotion detection to promote originality and creativity. We aim to unlock personality features with unmatched precision by working on the facial expressions, acknowledging that words cannot truly capture the essence of a person's emotions and personality. Our approach is a major step towards a new era in emotion detection, where a deeper understanding of human personalities may be gained through the combination of deep learning and image analysis. With the help of visual data, we can look into different sides of individuality and overstep the aspect emotion prediction with recommending the movies and songs.

1 Introduction

One of the most popular ways for people to convey their attitudes or feelings and thoughts is through their facial expressions. This makes it possible to predict someone's thoughts and feelings in another way. This project's main objective is to identify a person's emotions using transfer learning models like VGG16 and VGG19 and deep learning models and recommending movies and songs to make them better. Due to advancements in technology, multiple type of data is available. Data is also widely considered one of the most important resources for running automated systems Sonawane and Sharma (2020). Over the past few decades, technological innovation has been intertwined with any human effort in this context. A person's conduct and traits in a variety of contexts combine to form their emotions. Among other factors, a person's preference for websites, books, music, and movies can be influenced by their emotions or the emotions can also be influenced by the music or movies. In addition, emotions affect how one interacts with people and the surroundings. Personality testing is useful in a variety of contexts, including hiring, health advise, relationship guidance, career guidance. The conceptual foundation of emotional traits Suhasini and Badugu (2018), network behaviour, and personality relevance—which forms the basis of emotions detection—comes from the fact that people's behaviour is constant across time. Cross-domain consistency is necessary for building this model of emotions detection based on the theory Because emotion predicts behaviour consistently, it's a great tool for predicting individual behaviour. People with different emotions at different times, views toward social networks, and ways of using the route are all different. Although a person's emotions might vary greatly, even completely, depending on the type of network or the movies and songs they use, emotion detection still has a conceptual basis even when people use different networks in real lifeHealy et al. (2018). The relationship between network properties and personality projections became muddled as a result of the evolution of the network. According to research, a person's emotional characteristics, which are a psychological construct, can help to identify a person's behaviour and the songs and movies can help to change the emotion at a certain level. This presumption leads to a wide variety of personality models. Understanding people's emotions gives you a clue to what they like and don't like in different contexts.

1.1 Motivation and Project Background

Its hard to understand emotions as a great number of personality like performer,giver, counsellors, provider,etc; exists. By just reading a text or listening someone expressing their thoughts, we are unable to completely understand an individual's emotions. emotion detection via images would be the solution for the existing problem and helping to change the emotion in a better way via songs and movies. Emotion detection via images would help to identify will personalities or emotions even in any possible authorized settings. We are using transfer learning models to predict the emotions of images for this project. In this work, such as VGG16 and VGG19, transfer learning models can be applied. By improving the intelligence, security and efficiency of testing and education, it will help to increase performance. Therefore, the development of a model that uses images and transfer learning models in order to predict human emotion and making them better by recommending movies and songs is an objective of this research.
This study of research Jaiswal et al. (2020) refers to the research of human emotion detection from images. A study on an AI system will be able to detect emotions by facial expressions is presented in this study. It includes three steps: face detection, features extraction and emotion classification .
The study employs a CNN-based deep learning architecture for emotion detection. The model's results using two datasets: the facial emotion recognition challenge (FERC-2013) and the Japanese female facial emotion (JAFFE). Model accuracies are 70.14 and 98.65 percent for the FERC-2013 and JAFFE datasets, respectively. The subject of automatic facial expression analysis Shan et al. (2009) is both fascinating and difficult yet interesting, with notable implications for fields such as data-driven animation and human-computer interaction. Creating an accurate image of the faces as they are from source images is a major aspect of effective identification of emotion. In the research, they assess person-independent facial expression identification using facial representation based on statistical local features, Local Binary Patterns. Multiple databases are being used to cross-verify different machine learning techniques. The characteristics of LBP

for facial emotion recognition are demonstrated to be effective and efficient in multiple tests. We continue developing BoostedLBP to extract the most disproportionately high concentrations of LBCP features. Support for vector machine classification that uses BoostingLBP feature results in the best recognition performance.

1.2 Research Question

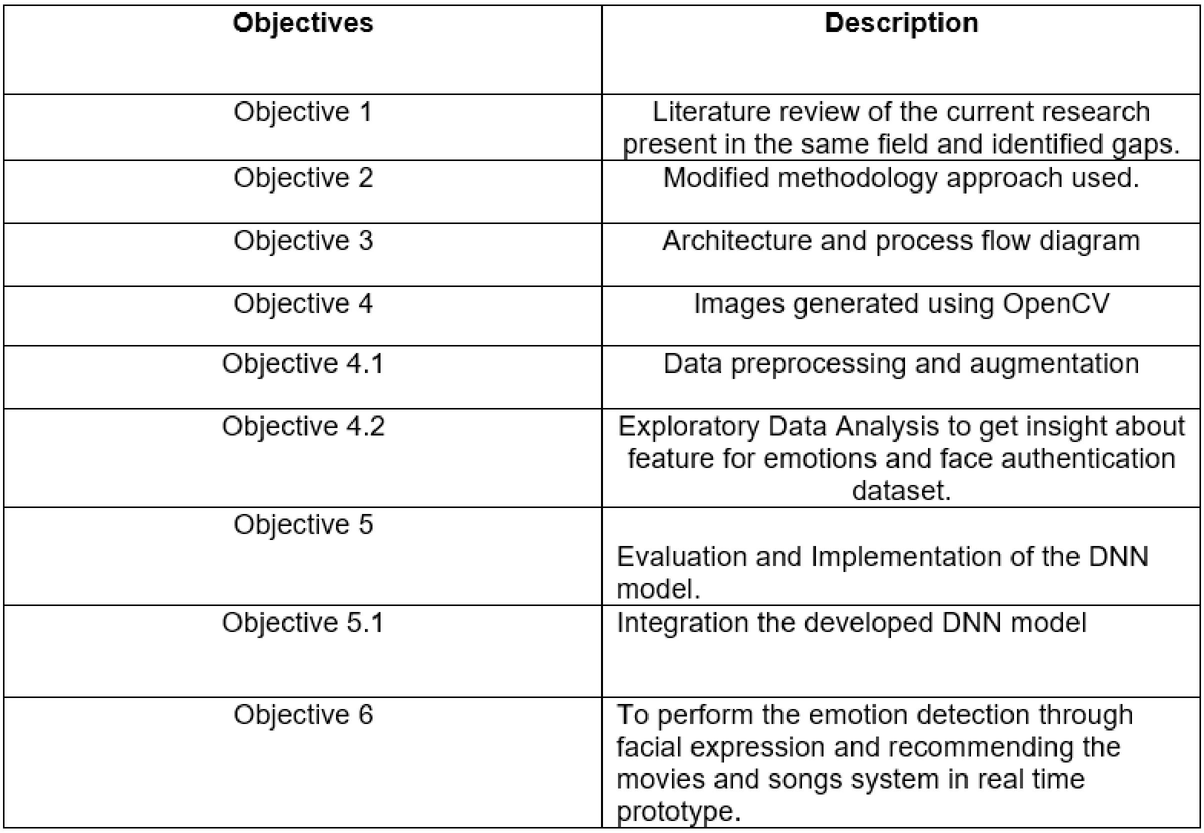
RQ: 'How far can OpenCV-generated photos be used to precisely train deep neural network architecture used to determine a person's emotions?'

Sub RQ: : 'How precise would the prediction of emotions by transfer learning models in this research perform in classifying emotions to understand a person's personality and recommending movies or songs while deploying it in real time as a web UI?'

In this experiment, deep neural network algorithms and their integration into a web interface for immediate use have been used to improve the accuracy and sensitivity of emotion detection.

1.3 Research Objectives

The research project's goal at each stage of development is shown in figure 1. All the steps are distributed in different phases is shown. From the collection of data to implementing the DNN models for the accuracy of models.



Objectives	Description
Objective 1	Literature review of the current research present in the same field and identified gaps.
Objective 2	Modified methodology approach used.
Objective 3	Architecture and process flow diagram
Objective 4	Images generated using OpenCV
Objective 4.1	Data preprocessing and augmentation
Objective 4.2	Exploratory Data Analysis to get insight about feature for emotions and face authentication dataset.
Objective 5	Evaluation and Implementation of the DNN model.
Objective 5.1	Integration the developed DNN model
Objective 6	To perform the emotion detection through facial expression and recommending the movies and songs system in real time prototype.

Figure 1: Project Management Objective

2 Related Work

Numerous scholars have made contributions to the field of Emotion Detection. While some have used machine learning techniques like CNN to predict outcomes, others have used psychological examinations to define emotional classifications. A synopsis of the literature review is provided below.

In this paperMaurya and Sharma (2022), dataset has been colletect through various social media platform. convolution neural network with keras has been used for face recognition. This model is trained with 24176 image sample and validated with 3006 image sample. Also, as this model has used dataset image of 48*48 size(grayscale), the model has been validated as it is able to detect all 5 emotions like happy, sad, neutral, angry and surprise. The model is set to 25 epochs and training accuracy is 69 percentage Singh et al. (2022). Also, the output is given as blue box around the face with the emotion detected as per the expression of the human.

This publication LokeshNaik et al. (2023) uses CNN to detect the emotion of a person based on their expression. CNN is one of the deep learning technique used to build a high prediction accuracy model.The image dataset is used to train the machine to recognize seven distinct facial expressionsM et al. (2023). The model's training and validation accuracy are comparable, indicating that it has the best fit possible and can generalize to the data. The model's accuracy has been tested to be between 70 percentage and 80 percentage, and it employs an Adamoptimizer to reduce the loss function.

In this research Reney and Tripathi (2015), Voila Jones Face detection method is used for detecting the face in input image. It is the most well-known and established face algorithm for image-based face detection. The Viola-Jones algorithm's fundamental idea is to scan a sub-window across an input image that can identify faces. The most crucial component for comparing face traits and sound Mel frequency components in any face and emotion identification system is the database. Features of the face are computed and stored in the database during the creation process. Then, using various algorithms, this database is used to evaluate the face and emotion. For evaluation of the face and detection of the emotion, KNN classifier is used. Efficiency of this prototype is 94.5 to 97 percentage.

The main aim of this research Bittal et al. (2023) is to maintain the attendance of students effectively. A multifarious face attendance system is developed using MT-CNN; which is excellent for its robustness, high accuracy and efficiency and VGGFace2 for large scale face recognition dataset model. This system's primary objective is to use VGGFace2 to identify several faces and map attendance based on that.
In order for us to examine the participants in the function or program. After thenGupta (2018), the faces that have been found can be through cross-referencing with the database of student faces. Attendance and participant records can be efficiently sustained with the aid of this useful method. Overall, the proposed system has the potential to improve the overall efficiency and effectiveness of educational attendance tracking institutions.

In order to identify local emotional importance in the fundamental frequency, this paper Arias et al. (2014) suggests using neutral reference models. The goal of this unique

approach, which is based on functional data analysis (FDA), is to capture the inherent diversity of F0 contours. The testing F0 contour is defined by how it is projected onto a basis of functions that represents the neutral models. We estimate the functional principal component analysis (PCA) projections for a given F0 contour, which are utilized as features in the emotion recognition process.Lexico-dependent (one functional PCA basis per sentence) and lexiconindependent (one functional PCA basis per sentence) models are used to assess the method.
The results of the experiment demonstrate that the suggested system can achieve binary emotion classification accuracy of up to 75.8 percentage, which is 6.2 percentage higher than that of a benchmark system trained with global F0 statistics. The method makes it easier to identify the specific emotional content that is being expressed within a sentence by applying it at the sub-sentence level (e.g., 0.5 s segments). The spontaneous corpus SEMAINE database is used to validate the method. The outcomes show that the suggested method for identifying emotional speechHua et al. (2006) in real-world applications can be used successfully.

Facial expression-based automatic emotion recognition is an interesting sub5 ject of study Mellouk and Handouzi (2020) that has been presented and used in a variety of contexts, including human-machine interfaces, safety, and health. In order to improve computer predictions, researchers in this discipline are interested in creating methods for interpreting, coding, and extracting facial expressions. Given deep learning's extraordinary success, its various architectures are being used to their full potential in order to improve performance.
This paper aims to investigate current advances in deep learning-based automatic facial emotion recognition (FER). We highlight the contributions addressed, the architecture, and the databases that were used. We also illustrate the progress that has been made by comparing the suggested approaches with the outcomes that were achieved. This study aims to assist and direct researchers by reviewing current literature and offering perspectives to advance this topic.

Human emotion is most naturally conveyed through textual language. Textual emotion recognition (TER) has gained significant attention in natural language processing because of its substantial academic and commercial possibilities. In recent years, TER has been heavily promoted and has gained increasing attention because to the enhanced development of deep learning technology. This paper Deng and Ren (2023) offers an organized overview of the most recent developments in TER, with a particular emphasis on methods utilizing deep neural networks.
TER methods are reviewed on word embedding, architecture, and training levels, respectively, based on how deep learning functions at each stage. Four issues were covered in our discussion of the remaining prospects and challenges: the lack of comprehensive and high-quality datasets; unclear emotional boundaries; partial extractable emotional information in texts; and TER in conversation. An organized and comprehensive review of deep TER technologies is produced by this article. It gives pertinent researchers the information they need to better grasp the current state of the field's research, lingering problems, and potential future directions.

Deep neural networks are increasingly being used to learn selective representations for automatic facial expression recognition (FER), as a result of the field's Li and Deng

(2022) recent success with deep learning techniques and the shift of FER from controlled laboratory settings to difficult in-the-wild scenarios. Current deep FER systems often concentrate on two key problems: expression-unrelated variables (such illumination, head posture, and identification bias) and overfitting resulting from insufficient training data. We present a thorough analysis of deep FER in this survey, along with datasets and techniques that shed light on these fundamental issues.
Firstly, we present the publicly available datasets that are often utilized in the literature together with established guidelines for their evaluation and data selection. Next, introduce existing unique deep neural networks and related training procedures for the state-of-the-art in deep FER . These networks have been created for the FER project by combining Static Images with Dynamic Images, and we are analyzing their advantages and disadvantages. A summary is also provided for experimental comparisons and competitive results using commonly utilized benchmarks.

One of the most significant aspects of being human is emotion. Without the capacity to interpret emotions, robots and computers are unable to interact with people in a natural way. This work Murugappan (2011) introduced the use of Electroencephalogram (EEG) data for the classification of human emotions. Twenty volunteers' EEG signals are gathered using 62 active electrodes positioned throughout the scalp in accordance with the International 10-10 standard. A methodology based on audio-visual stimuli—video clips—has been developed for generating the distinct emotions. The Surface Laplacian filtering method is used to preprocess the raw EEG data, and the Wavelet Transform (WT) is used to break them down into five distinct EEG frequency bands: delta, theta, alpha, beta, and gamma. In order to obtain features of statistics from a processed signal, three different wavelet functions were investigated namely 'db4','db8, 'sym8' and 'coif5'. In this work, the effectiveness has been assessed for classifying emotions on two different EEG channels 62 and 24. In order to validate statistical data, linear nonlinear K Nearest Neighbor Classifiers are used for classification using five fold cross validation. The maximum average classification rate that KNN can provide is 82.87 percentage on 24 channels and 82.87 percentage on 62 channels. Lastly, we demonstrate the KNN's average and individual classification accuracy to support the effectiveness of our emotion identification system.

The issue of facial emotion detection has attracted a great deal of attention over the last few years. In order to recognise facial expressions, many methods have been developed. They describe a deep learning model for facial expression classification in this paper Lalitha et al. (2021). The model recognizes seven expressions in particular: joyful, sad, angry, fear, disgusted, surprised and neutral. Haar cascade classifiers were used for face detection. The Facial Expression Recognition 2013 [FER2013] dataset was used to train the model, and the outcomes were examined. The proposed model performed wellKishan Kondaveeti and Vishal Goud (2020), with training accuracy of 78 percentage and validation accuracy of 67 percentage. The system can be used for both static photographs and realtime video.

3 Methodology

3.1 Introduction

To understand the Research the technique of Emotion Detection, it is necessary to explain the three main elements of the technique. It is important to understand the facts, and this cannot be emphasized enough. This section describes the data that are used to train a model for identifying emotions. Emotion detection utilizing 7 data-driven emotion information and facial recognition produced programmatically with the OpenCV and from kaggle. Before the data is utilized to train the transfer learning models like VGG16 and deep learning models, it must first be pre-processed, transformed, and improved.
To get a better idea of how the emotion detection web application prototype works together with transfer learning models to create a system that is capable of understanding human emotions. There are subsections for each of these steps. KDD was used in this study to control the project management and development phases of a difficult process for identifying interesting patterns, real as well as valuable ones from an image collection gathered during these studies; together with key decision making.
In general this involves several steps, e.g. choosing an appropriate method of data mining to extract patterns for classification from the image dataset, preprocessing and transforming filesfrom a random set of images in order to create and select subsets using OpenCV. After the completion of each step, the model is tested, optimized, assessed, and made available for usage in real time.

3.2 Data Collection

3.2.1 Facial EMotion Dataset

The OpenCV framework is used to generate the various emotion data sets that are needed to train the DNN model to carry out specific functions inside the system, such as emotion detection. To make specific images based on face expression movement, an input video source from the camera is accessed and processed. It draws 100 images representing each emotion out of the background during one code execution using wait key 1. Open a display window, and the process of taking pictures begins.
The primary location of the emotional data set will be the directory where the path has been assigned. Using the labels that were assigned to each one, such as "Angry", "Disgust", "Fear", "Happiness", "Neutral", "Sad" and "Surprise" seven directories were made.

3.2.2 Face Authentication Dataset

In this data collection, more than 2000 facial images were generated programmatically with openCV and 35000 facial images taken from kaggle. On every picture is the name of the person it features. Each person pictured is represented by 100 distinct photographs in the data set of 7 different emotions of the collection.
The model that will be used to apply facial expression recognition for system authentication will be trained using this dataset. The created DNN model assesses many facial features, like the width of your nose and the location of your eyes, and combines all of this data into a single code that serves as both an identity and a way of identification.

3.3 Data Pre-processing

To improve the quality of the data generated programmatically in neural networks and to provide a better estimation of realtime models without any loss of classification, preprocessing images are used. Some of the methods used in this research to improve neural network inputs are Interpolation, rescaling, dilation, opening, closing, and determining whether the data format, such as the RGB channel, is first or last and feeding the input shape correctly. All preprocessing is done with the Keras API utilizing built-in parameters.

3.4 Data Augmentation

To increase and diversify the datasets generated as well as to block neural networks from learning unrelated features,we performed data augmentation. Side-on transformation techniques are used because the model is an emotion detection system and will be merged in real-time.

When training the proposed transfer learning network and the custom DNN for feature extraction, the learning rate is maintained constant such that the change in accuracy is only attributable to augmentation . Every time we use augmentation to train our network, a new face authentication image and every change in the expression is generated for each epoch.

In this way, even if there is a fresh copy of these photographs in each epoch, the model sees an identical number of photos as it did with the first training set. In consequence, a growing amount of images have been seen in the model for each epoch.

3.5 Data Exploratory Analysis

The OpenCV framework is utilized to automatically produce expression and facial authentication datasets, which are plotted below as of photos. 20 frames per second were required and half of the data is taken from kaggle.

Some of the dataset's looks are shown in the pictures below along with their corresponding class names.

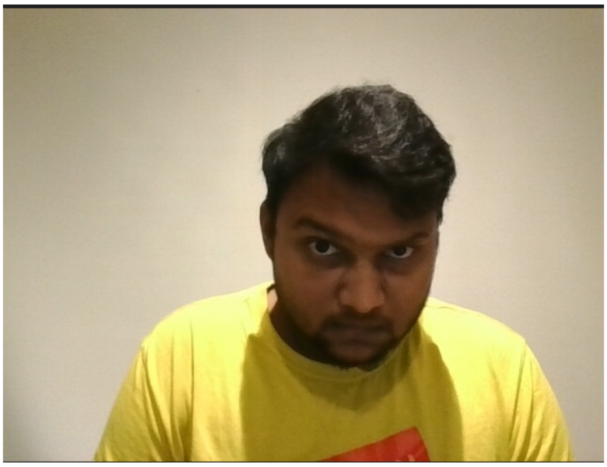


Figure 2: Emotion - Angry

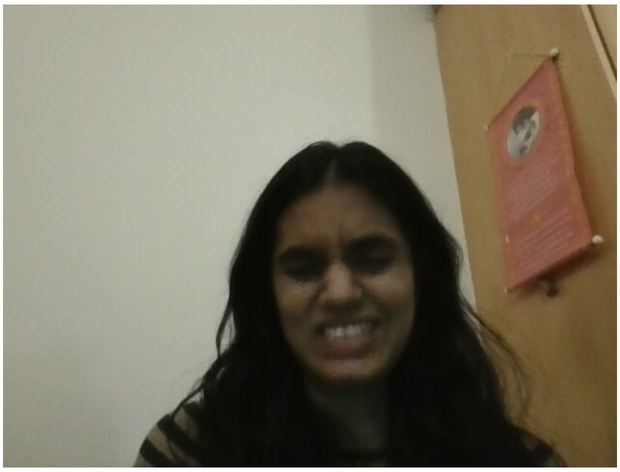


Figure 3: Emotion - Disgust

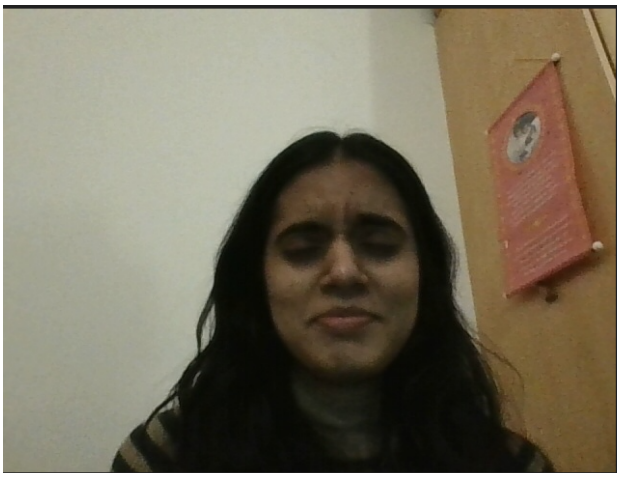


Figure 4: Emotion - Fear

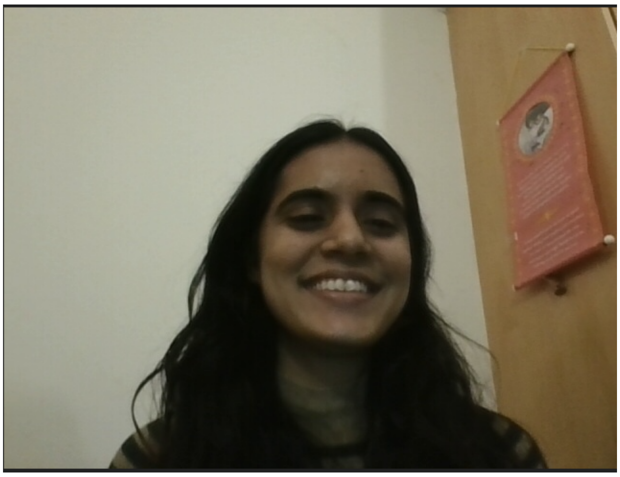


Figure 5: Emotion - Happy

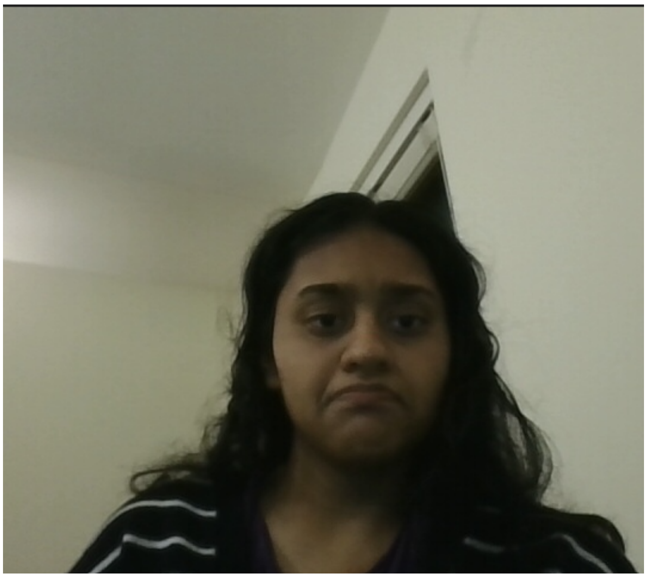


Figure 6: Emotion - Sad

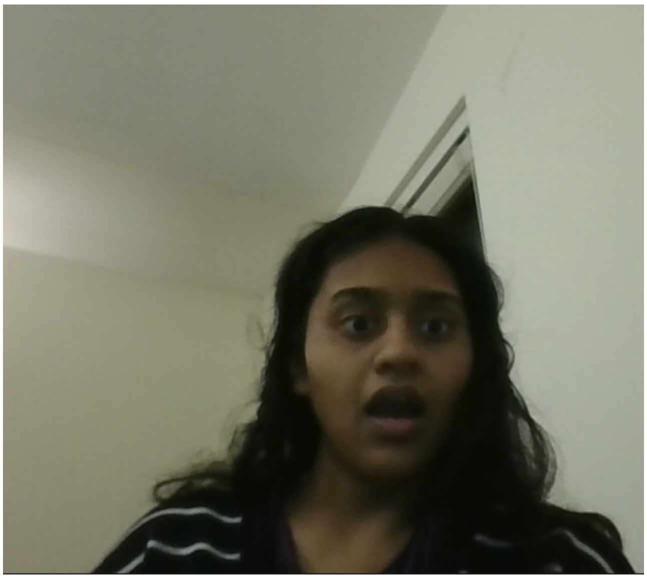


Figure 7: Emotion - Surprise

The below in Figure 8 displays the color distribution and brightness of each primary feature in the image, such as RGB, because the image is a three channel image with a resolution of 244x244 pixels .
As you advance in pixels in a rightward direction, the color becomes more intense and vibrant. The below image showing the analysis and feature reading of the facial expression.

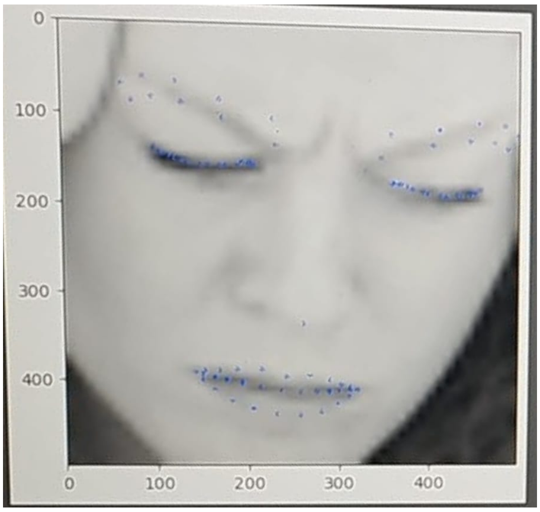


Figure 8: Visualization of detecting facial expressions

4 Design Specification

The following frameworks and technologies shown in the image 9 are used in order to execute the suggested design and make the emotion detection system prototype as a web application to monitor the created model's performance in real-time. The suggested design would be implemented using Keras, a deep learning application interface built on Python that serves as the front-end of the deep learning platform TensorFlow. This enables the quick realization of the emotion detection system concept, which is essential for effective research.

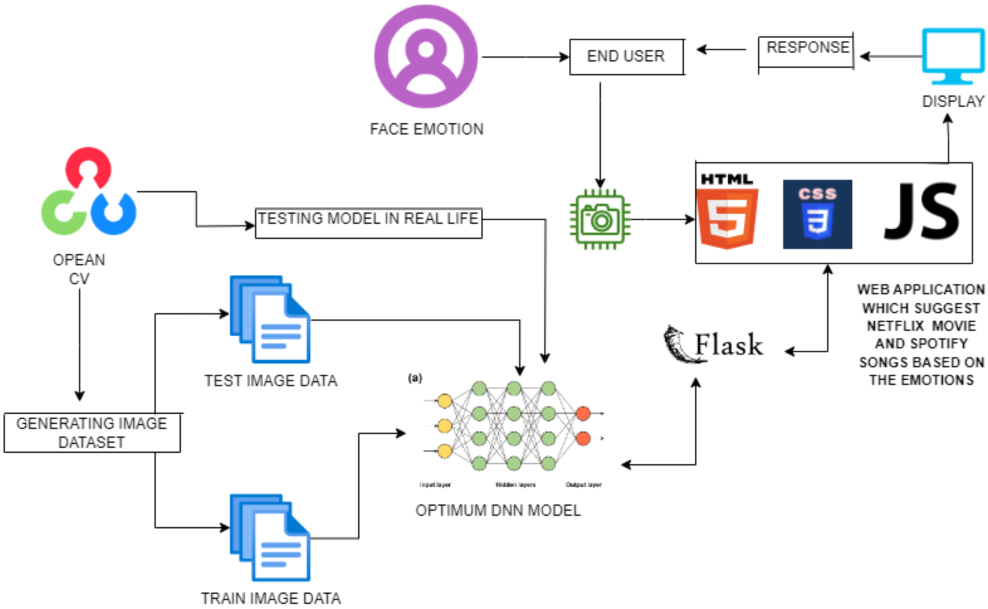
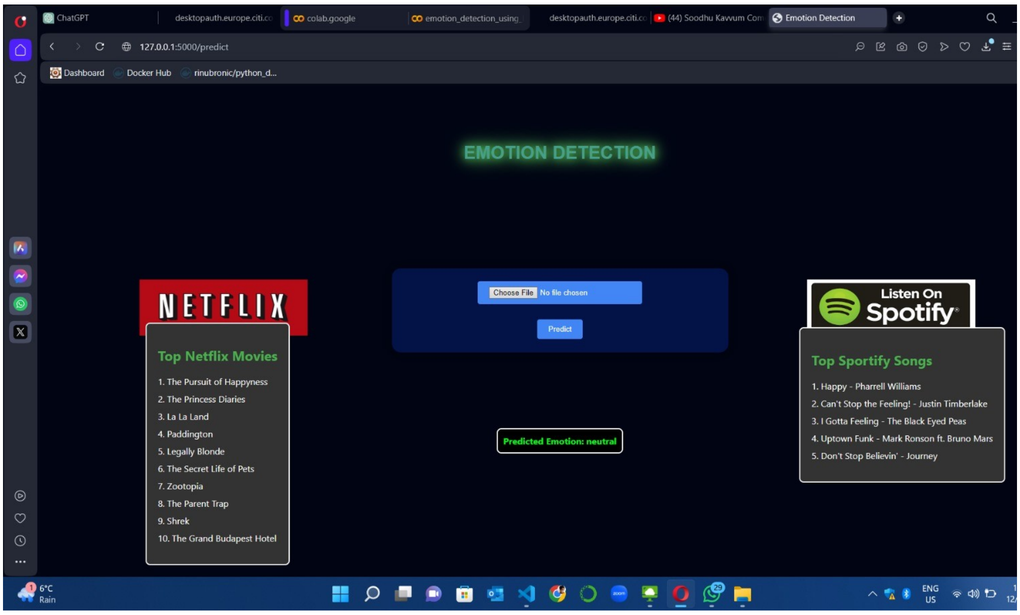


Figure 9: Design Specification

An open source deep learning framework called TensorFlow-2 efficiently performs low

level tensor operations on the CPU and GPU during the whole operation. The Emotion detection prototype is built using the flask web framework, a Python programming language, which combines the front end with an exporting deep learning model.for the frontend we have used HTML and backend implementation is done by python flask framework. For styling the application we used CSS. In the front-end application, there will be an image upload section where we can upload the picture.
After that it will predict the image and it will show the current emotion of the picture. According to the emotion the output will be generated. In the left side of the application it will suggest the Netflix movies to watch and the right side of the application it will suggest the Spotify music for the user. The OpenCV module gives you access to your device's camera or to choose a pic from dataset and the ability to do any computer vision operations.



Top Netflix Movies

1. The Pursuit of Happyness
2. The Princess Diaries
3. La La Land
4. Paddington
5. Legally Blonde
6. The Secret Life of Pets
7. Zootopia
8. The Parent Trap
9. Shrek
10. The Grand Budapest Hotel

Predicted Emotion: neutral

Top Sportify Songs

1. Happy - Pharrell Williams
2. Can't Stop the Feeling! - Justin Timberlake
3. I Gotta Feeling - The Black Eyed Peas
4. Uptown Funk - Mark Ronson ft. Bruno Mars
5. Don't Stop Believin' - Journey

Figure 10: Front-end

5 Implementation

Phase 1: The OpenCV framework in Python is used to frame and build the face expression image dataset. The laptop's built-in camera is accessible via the open CV framework. Quickly, one hundred pictures were taken and saved along the designated path.
Seven different feelings picture datasets containing more than 2000 images were produced by repeating this technique. Each of these photos was categorized into a separate folder based on a different sentiment. Half of the face images data-set is taken from kaggle.

Phase 2: In order to retrieve these photographs programmatically in Google Colab Pro, they were all pushed to Google Drive. Google Colab Pro is used for this project development phase and exporting an optimal model to connect it with the web application because it is an excellent tool for deep learning jobs.
With its great version that gives users access to Google processing resources like GPUs

and TPUs to train and develop deep learning models quicker using massive picture datasets, it's a hosted Jupyter notebook that requires no installation.

Phase 3: After accessing the image datasets, a python framework called TensorFlow is utilized which holds libraries to develop multi-layer large-scale neural networks and to perform image data pre-processing, data augmentation. Three models such as CNN, Custom CNN and VGG16 were built to extract features from the images and classify the Emotions accordingly to perform specific tasks in the prototype.
All these models were evaluated using different metrics such as Accuracy, Precision, Recall, Validation Accuracy, Validation Precision, Validation Recall. The optimum model is exported in the 'h5' file format for integration.

Phase 4: The front end of the web application is designed using HTML, CSS, and Javascript. To access the device camera a separate class in javascript is implemented. The integration between the front end and the exported optimum deep learning model is established using the Python Flask framework which is clearly explained in the Figure.

Phase 5: After integrating the exported optimum deep learning model with the front end. The web application server is hosted and UAT is performed to check the real-time execution of the model. Initially, face authentication is performed to maintain the individuality and then to understand the features of the facial expressions.
The javascript class for accessing the device camera is programmed to capture a frames per second. After capturing the images, the flask framework sends the request for a prediction to the machine learning model. The result from the model is sent back to the front end as a response to perform the recommendations of the movies or songs.

5.1 Implementing diffractive deep neural network for Emotion detection

Once required data has been produced to train and evaluate the transfer learning and custom CNN models, including VGG16. We perform an explanatory analysis on the dataset. Prior to developing the model, thoroughly comprehend the image dataset. Figure illustrates how the data is pre-processed by scaling and altering each image to the same desired size.
The image is converted to any constant size of for training, the custom CNN model, and for training and validating transfer learning models like VGG16. After that, the image is enhanced utilizing various factors, including as "width shift range," "height", "zoom range", "rescale", "rotation range" and "horizontal flip" 'shift range' and the dataset's image is enlarged.
Following the construction of all the models,the effectiveness of every DNN model utilizing several variables and parameters, including the number of epochs, dataset size, DNN model run duration, accuracy, and precision, recall accuracy, and recall of validation. Lastly, the instance in which It also takes into account the processing unit on which the model is trained.

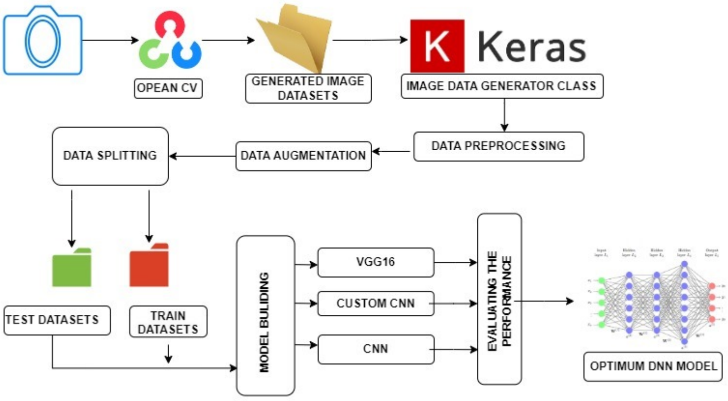


Figure 11: Selection of the Model

5.1.1 Custom CNN diffractive DNN architecture for emotion detection

The customized CNN architecture shown in the Figure is designed such that each Conv2D and MaxPooling2D layer produces a 3D tensor of shape with channels, width, and height. Because it produces images, it has three channels. Prior to being added to the input layer, each image in the dataset is resized to the desired size. The width and height measures get smaller as one moves further within the network.The number of output channels for each Conv2D layer is controlled by units and the activation layer of the final dense layer. Using the Softmax activation function, which works better for polynomial classifications with multiple polynomials, the last dense layer in this architecture contains five units that can fire.

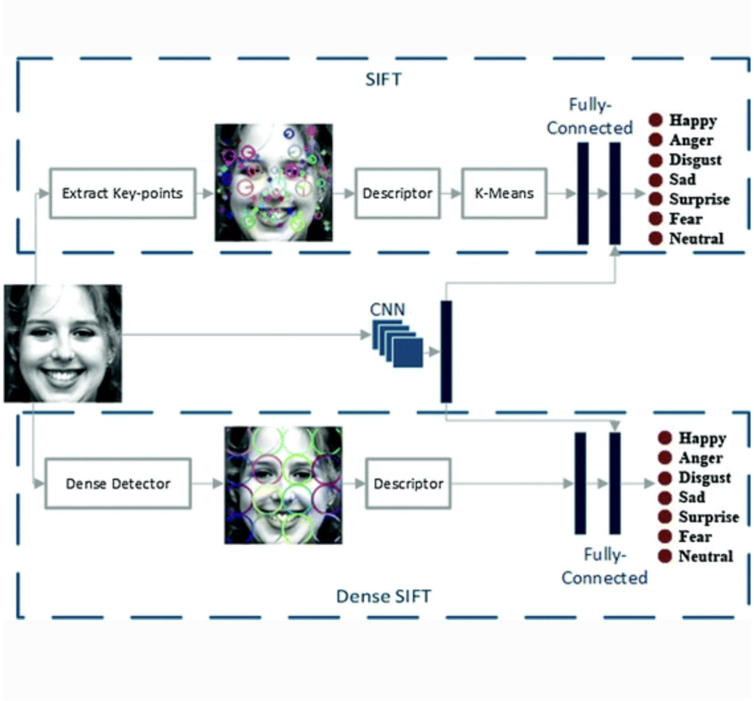


Figure 12: CNN Architecture

5.1.2 VGG16 Transfer learning DNN architecture for emotion detection

A two-dimensional image (224 × 224 x 3) seen in figure serves as the network's input. The first two layers include 64 channels with a 3*3 filter size and the same padding. Afterwards, a max pool layer (3 x 3) is added, followed by two layers of convolution layers with 256 filter size and filtering size. A stride (2 x 2) maxpooling layer, which is the same as the preceding layer, comes next. The next two convolution layers have filter sizes of 3 and 3, along with 256 filters. Next, there are a max pool layer and two sets of three convolution layers. Every filter has 512 identically sized (3 x 3) filters with the same padding. After that, a two-layer convolution stack receives this image. In order to modify the amount of input channels, it additionally uses 1 x 1 pixels in certain of the layers. As seen in the Figure, 1-pixel padding is used after each convolution layer to preserve the spatial information of the image.

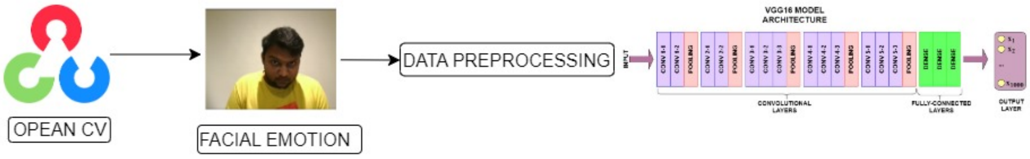


Figure 13: VGG16 Architecture

5.1.3 Feature extraction and engineering

The DNN network used in this work enables a picture to extract the activations at the final maxpooling layer before the completely linked layers. The maxpooling layer's output, which we flatten into a feature vector with a dim of 21,055, is shown in volume form.
A set of photographs undergo this process for every picture, yielding 2000 x 21,055-dim feature vectors in total.

5.2 Implementing diffractive Transfer learning model for Facial Authentication

The usage of face authentication helps to comprehend the emotions. In order to achieve facial authentication in the first place, use of transfer learning, a deep learning model was constructed and compared in order to determine which model performed best overall. OpenCV was used to create the person's facial data set. After processing, these datasets were divided into training and testing sets. Thus, where the weights are retrieved from ImageNet, such as VGG16 were employed. Our Facial Emotion Recognition system is highly accurate and versatile due to its feature extraction and careful engineering methods. VGG16 is a variant of the VGG model with sixteen layers: three completely linked levels, five MaxPool layers, one SoftMax layer for classification, and sixteen weight layers, including thirteen convolutional layers. The other models employed in this study, including VGG16

5.2.1 Feature extraction and engineering for Facial Authentication model

In order to train the transfer learning models and authenticate the prototype, all of the facial images created with OpenCV are transmitted to the final max-pooling layer before to the fully connected layers, where activations are taken.
The results of the max-pooling layer in volume form, which we then flatten to a feature vector of dimension 21,055. Using this method, 1500 x 21,055 dimensional feature vectors are produced from A set of 2000 pictures.

6 Evaluation

6.1 Evaluating the performance of the diffractive deep neural network

One of the main validation techniques used in this study to train the CNN and transfer learning models like VGG16 is cross-validation. It offers an accurate evaluation of a model's effectiveness with hypothetical data. Two subsets of the dataset are used for training and evaluating the models' performance: the held-out emotion dataset is used to test the models on all but one of the subsets.
Until the held-out emotion set can be chosen from each of the subgroups depicted in the picture, the process is repeated. The performance metric is then aggregated over all of the models that have been built. Numerous metrics, including accuracy, precision, recall, validation accuracy, validation precision, and validation recall, were used in this study. When selecting the final model for integration, other factors such as the number of epochs, processing unit, data size, and training duration are also taken into account. Following the export of the best model and a comparison of all model performances

Figure 14: Picture taken with OpenCV

6.1.1 Analyzing Custom CNN diffractive DNN architecture for Emotion detection

During the 123 MB of data used to train the unique CNN model, which consists of three convolution layers. Figures makes this very evident: it took 190 seconds to train the model using 20 epochs. The model was ran on a Google Colab Pro in order to access a faster GPU. The CNN-1 models display 51.89 and CNN2-2 model displays 88.32, precision, and recall, respectively.
A few other parameters, including validation recall, accuracy, and precision —which are the most crucial metrics indicating the model's real-time performance—were also computed. The loss and accuracy for each epoch of the model's training are displayed in the line graph below.

CNN 1

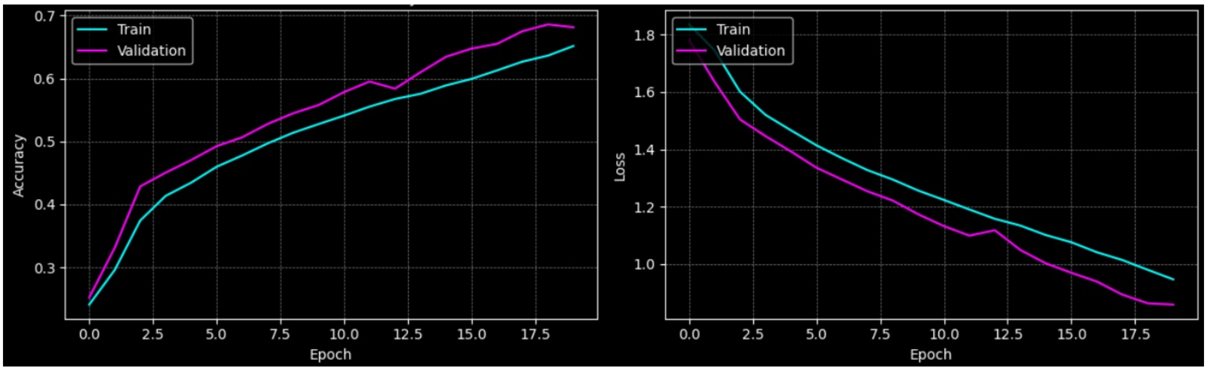


Figure 15: validation accuracy and validation loss for CNN 1

CNN 2 with adding more layers.

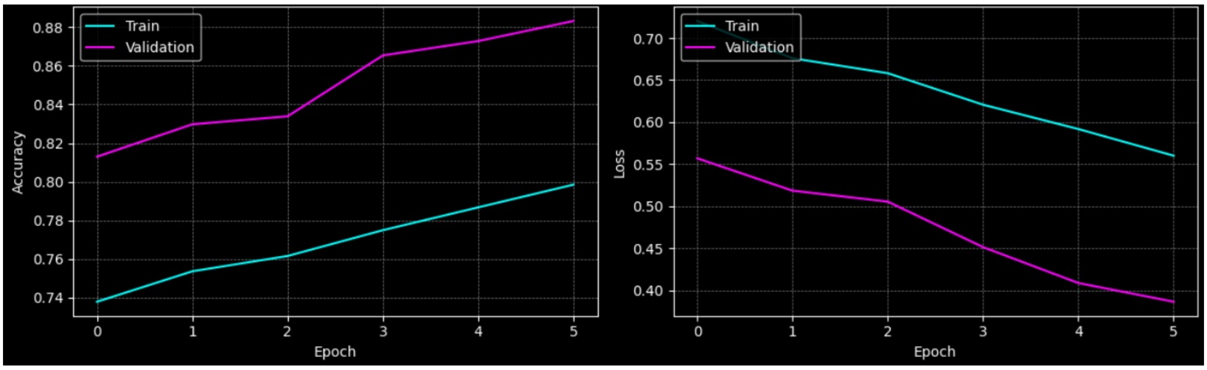
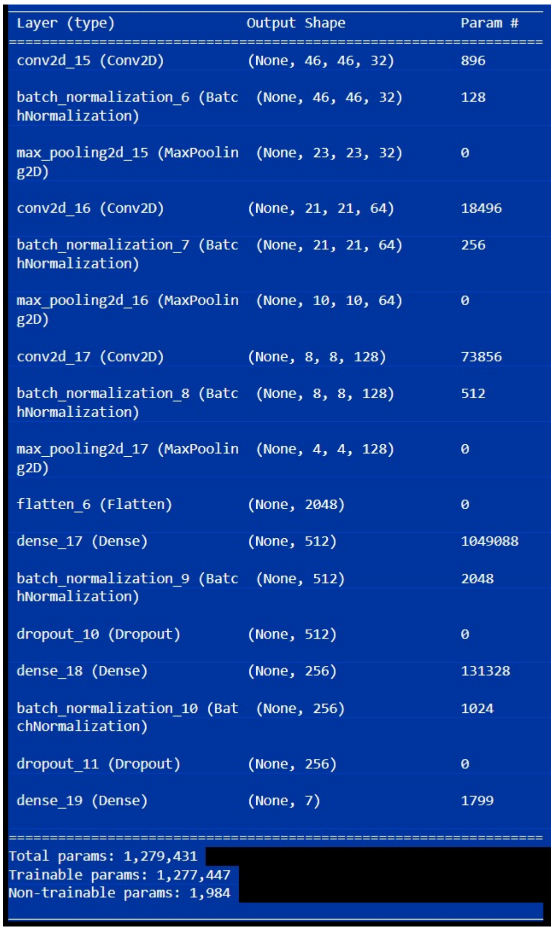


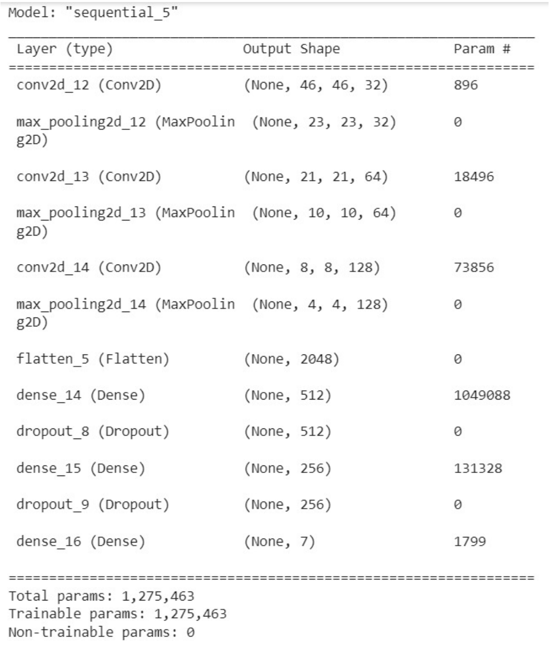
Figure 16: validation accuracy and validation loss for CNN 2

Fine-tuning the model: The implementation of call back, early stopping, and reduceLROnPlateau helps reduce saturation and overfitting of the model parameters. Certain factors, like a slower learning rate, can be used to the model's advantage when learning remains stationary. The learning rate is decreased by a factor of 2–10. If, after a certain number of "patience" epochs, no improvement is observed, this callback analyzes a quantity and lowers the learning rate.



```
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_15 (Conv2D)          (None, 46, 46, 32)        896

 batch_normalization_6 (Batc  (None, 46, 46, 32)       128
 hNormalization)

 max_pooling2d_15 (MaxPoolin  (None, 23, 23, 32)       0
 g2D)

 conv2d_16 (Conv2D)          (None, 21, 21, 64)        18496

 batch_normalization_7 (Batc  (None, 21, 21, 64)       256
 hNormalization)

 max_pooling2d_16 (MaxPoolin  (None, 10, 10, 64)       0
 g2D)

 conv2d_17 (Conv2D)          (None, 8, 8, 128)         73856

 batch_normalization_8 (Batc  (None, 8, 8, 128)        512
 hNormalization)

 max_pooling2d_17 (MaxPoolin  (None, 4, 4, 128)        0
 g2D)

 flatten_6 (Flatten)         (None, 2048)              0

 dense_17 (Dense)            (None, 512)               1049088

 batch_normalization_9 (Batc  (None, 512)              2048
 hNormalization)

 dropout_10 (Dropout)        (None, 512)               0

 dense_18 (Dense)            (None, 256)               131328

 batch_normalization_10 (Bat  (None, 256)              1024
 chNormalization)

 dropout_11 (Dropout)        (None, 256)               0

 dense_19 (Dense)            (None, 7)                 1799

=================================================================
Total params: 1,279,431
Trainable params: 1,277,447
Non-trainable params: 1,984
```

Figure 17: CNN-1 Model structure



```
Model: "sequential_5"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_12 (Conv2D)          (None, 46, 46, 32)        896

 max_pooling2d_12 (MaxPoolin  (None, 23, 23, 32)       0
 g2D)

 conv2d_13 (Conv2D)          (None, 21, 21, 64)        18496

 max_pooling2d_13 (MaxPoolin  (None, 10, 10, 64)       0
 g2D)

 conv2d_14 (Conv2D)          (None, 8, 8, 128)         73856

 max_pooling2d_14 (MaxPoolin  (None, 4, 4, 128)        0
 g2D)

 flatten_5 (Flatten)         (None, 2048)              0

 dense_14 (Dense)            (None, 512)               1049088

 dropout_8 (Dropout)         (None, 512)               0

 dense_15 (Dense)            (None, 256)               131328

 dropout_9 (Dropout)         (None, 256)               0

 dense_16 (Dense)            (None, 7)                 1799

=================================================================
Total params: 1,275,463
Trainable params: 1,275,463
Non-trainable params: 0
```

Figure 18: CNN-2 Model structure

6.1.2 Evaluating VGG16 Transfer learning DNN architecture for emotion detection

The proprietary VGG16 transfer learning model, consisting of convolution and max pool layers, was trained uniformly throughout the architecture using 123 MB of data and 2000 photographs and the data fromkaggle. Figure clearly illustrates that training the model with 20 epochs takes 240 seconds. The Google Colab Pro was used to run the model in order to have access to a faster GPU. The models had 68.12 percent accuracy respectively. The most crucial characteristics that represent the model's performance in realtime were also calculated, including validation accuracy, precision, and recal. The line graph below shows the accuracy and loss for each training session of the model.

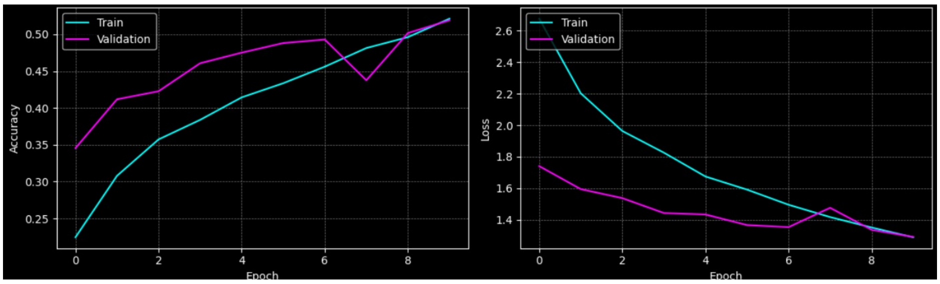
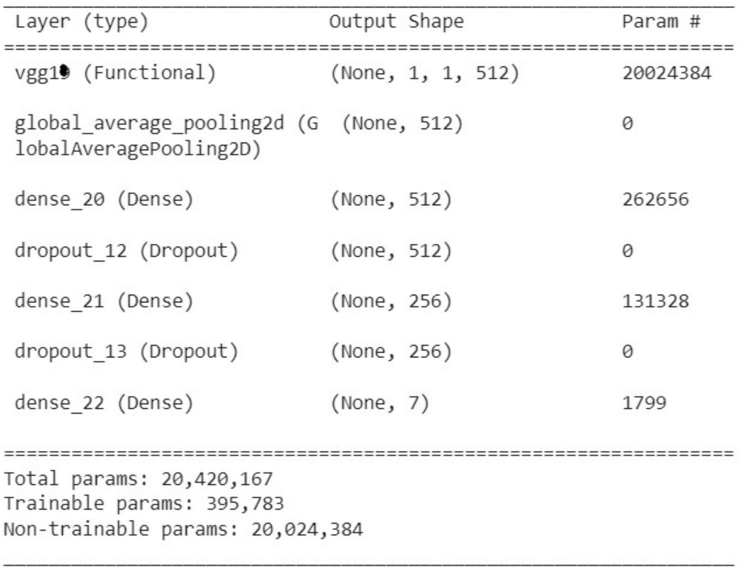


Figure 19: validation accuracy and validation loss for VGG16

Fine-tunning the model: The VGG-16 model was modified using image augmentation to see if this would improve model accuracy. Using the same VGG-16 model object that was kept in the transfer model variable from our previous model, unfreeze the fifth convolution block while keeping the first four blocks frozen.

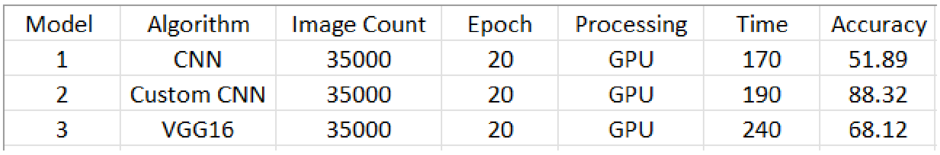


```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 vgg16 (Functional)          (None, 1, 1, 512)         20024384

 global_average_pooling2d (G  (None, 512)              0
 lobalAveragePooling2D)

 dense_20 (Dense)            (None, 512)               262656

 dropout_12 (Dropout)        (None, 512)               0

 dense_21 (Dense)            (None, 256)               131328

 dropout_13 (Dropout)        (None, 256)               0

 dense_22 (Dense)            (None, 7)                 1799

=================================================================
Total params: 20,420,167
Trainable params: 395,783
Non-trainable params: 20,024,384
_________________________________________________________________
```

Figure 20: VGG16 model structure

6.2 Comparison of Developed Models and Discussion

Following the building of the suggested model architecture with the Keras framework and model training with OpenCV-generated images for each model. To reach the best results, many parameters are adjusted for each CNN and Transfer learning model VGG16. A comparison of all the models in Figure is presented when compared to the above matrices and parameters. It is clear from the image that Custom CNN is performed way better than the rest two and the run time is also low.



Model	Algorithm	Image Count	Epoch	Processing	Time	Accuracy
1	CNN	35000	20	GPU	170	51.89
2	Custom CNN	35000	20	GPU	190	88.32
3	VGG16	35000	20	GPU	240	68.12

Figure 21: Metrics for training data

The Custom CNN model is then exported as h5 format and integrated with the dataset and openCV to take the images. The images will be loaded using the predicted class(emotions). Then the output will be getting in an array format for each emotion. The model is integrated by the from-end of the website of the prototype using the python flask framework.

7 Conclusion and Future Work

To accomplish the study and the system, different deep learning models such CNN, custom CNN, and additional transfer learning models like VGG16 were constructed. This fresh and creative human-machine interaction has a deep awareness of human emotions. This will make it easier to comprehend each person's personality through their feelings. This technology will be a significant breakthrough in this era of technology and will benefit society by understanding the human emotions and help them making it better. The factors listed in the sections above are used to determine the best model. In order to advance this research, many DNN models will be processed in parallel at the same time in the future rather than using a single model to identify emotions, and converting the prototype in the full fledged application. After comparing the output from each model, the response on the web application's front end will be displayed as the human emotions and accordingly recommending the songs and movies that the majority of the models were able to identify. This will be more accurate and totally prevent the validation loss.

Acknowledgment

I would want to thank the following individuals because without them, I could not have completed this research project or advanced this far in my master's program. The staff at the National College of Ireland, particularly my supervisor, Taimur Hafeez, helped me with my research by sharing his expertise and understanding of the subject. And a particular thank you to my housemates and friends, who inspire me to achieve academic success.

References

Arias, J. P., Busso, C. and Yoma, N. B. (2014). Shape-based modeling of the fundamental frequency contour for emotion detection in speech, *Computer Speech & Language* **28**(1): 278–294.

Bittal, V., Jagdale, V., Brahme, A., Deore, D. and Shinde, B. (2023). Multifarious face attendance system using machine learning and deep learning, *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 387–392.

Deng, J. and Ren, F. (2023). A survey of textual emotion recognition and its challenges, *IEEE Transactions on Affective Computing* **14**(1): 49–67.

Gupta, S. (2018). Facial emotion recognition in real-time and static images, *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pp. 553–560.

Healy, M., Donovan, R., Walsh, P. and Zheng, H. (2018). A machine learning emotion detection platform to support affective well being, *2018 IEEE International Conference on Bioinformatics and Bio-medicine (BIBM)*, IEEE, pp. 2694–2700.

Hua, A., Litman, D. J., Forbes-Riley, K., Rota-ru, M., Tetreault, J. and Purandare, A. (2006). Using system and user performance features to improve emotion de-tection in spoken tutoring dialogs, *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vol. 2, pp. 797–800.

Jaiswal, A., Krishnama Raju, A. and Deb, S. (2020). Facial emotion detection using deep learning, *2020 International Conference for Emerging Technology (INCET)*, pp. 1–5.

Kishan Kondaveeti, H. and Vishal Goud, M. (2020). Emotion detection using deep facial features, *2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI)*, pp. 1–8.

Lalitha, S. K., Aishwarya, J., Shivakumar, N., Srilekha, T. and Kartheek, G. C. R. (2021). A deep learning model for face expression detection, *2021 International Conference on Recent Trends on Electronics, Information, Communication Technology (RTEICT)*, pp. 647–650.

Li, S. and Deng, W. (2022). Deep facial expression recognition: A survey, *IEEE Transactions on Affective Computing* **13**(3): 1195–1215.

LokeshNaik, S., Punitha, A., Vijayakarthik, P., Kiran, A., Dhangar, A. N., Reddy, B. and Sudheeksha, M. (2023). Real time facial emotion recognition using deep learning and cnn, *2023 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–5.

M, S., G, E. R., M, N., K, J., V, N. and R, P. (2023). Detection and recognition of face using deep learning, *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, pp. 72–76.

Maurya, A. and Sharma, V. (2022). Facial emotion recognition using keras and cnn, *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 2539–2543.

Mellouk, W. and Handouzi, W. (2020). Facial emotion recognition using deep learning: review and insights, *Procedia Computer Science* **175**: 689–694.

Murugappan, M. (2011). Human emotion classification using wavelet transform and knn, *2011 International Conference on Pattern Analysis and Intelligence Robotics*, Vol. 1, pp. 148–153.

Reney, D. and Tripathi, N. (2015). An efficient method to face and emotion detection, *2015 Fifth International Conference on Communication Systems and Network Technologies*, pp. 493–497.

Shan, C., Gong, S. and McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study, *Image and vision Computing* **27**(6): 803–816.

Singh, S. K., Thakur, R. K., Kumar, S. and Anand, R. (2022). Deep learning and machine learning based facial emotion detection using cnn, *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 530–535.

Sonawane, B. and Sharma, P. (2020). Deep learning based approach of emotion detection and grading sys-tem, *Pattern Recognition and Image Analysis* **30**: 726–740.

Suhasini, M. and Badugu, S. (2018). Two step approach for emotion detection on twitter data, *International Journal of Computer Applications* **179**(53): 12–19.