

A REPORT ON  
**“EMOTION BASED MUSIC PLAYER”**

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE  
IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE AWARD OF THE DEGREE

OF

**BACHELOR OF ENGINEERING  
(Information Technology)**

SUBMITTED BY

<b>Mahima Baliyan</b>	<b>B150078502</b>
<b>Abhishek Karanjekar</b>	<b>B150078530</b>
<b>Jeevan Pawar</b>	<b>B150078543</b>
<b>Janhavi Wagh</b>	<b>B150078560</b>



**DEPARTMENT OF INFORMATION TECHNOLOGY**  
**PVG's COLLEGE OF ENGINEERING AND TECHNOLOGY**  
44, VIDYANAGARI, PARVATI, PUNE 411009  
**SAVITRIBAI PHULE PUNE UNIVERSITY**  
**2018 -2019**



## CERTIFICATE

This is to certify that the project report entitled

## EMOTION BASED MUSIC PLAYER

### SUBMITTED BY

Mahima Baliyan	B150078502
Abhishek Karanjekar	B150078530
Jeevan Pawar	B150078543
Janhavi Wagh	B150078560

Is a bonafide work carried out by them under the supervision of Prof. N.R. Sonawane and is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University for the award of Degree of Bachelor of Engineering (Information Technology). This project report has not been earlier submitted to any other institute or University for the award of any degree or diploma.

Internal Guide

Head of Department

Department of Information  
Technology

Department of Information  
Technology

External Examiner

Principal

Place :

PVG'S COET, Pune

Date :

## ACKNOWLEDGMENT

It gives us great pleasure in presenting the preliminary project report on “Emotion Based Music Player”.

We are highly indebted to our guide Prof. N. R. Sonawane and Head Of Department Prof. S. S. Dixit and our review panel, Prof. M.R. Apsangi and Prof. D. T. Varpe for their guidance and constant supervision. This project would not have been possible without the kind support and help of many individuals and organizations.

We would like to express our gratitude towards all the faculty members of IT department for their kind cooperation and encouragement which helped us in completion of this project. We would like to express our special gratitude and thanks to project coordinator for giving us such attention and time.

Our thanks and appreciations also goes to our colleagues in developing the project and people who have willingly helped us out with their abilities.

Mahima Baliyan  
Abhishek Karanjekar  
Jeevan Pawar  
Janhavi Wagh

## ABSTRACT

Music plays a very prominent role in an individual's life. Conventional method of listening to the music requires browsing through a playlist of songs for the current mood. Migrating the task to the computer vision technology enables automation and enhancing the user experience. Facial expressions are an important key factor in determining the emotion. This project proposes an efficient and accurate emotion based music player deployed as a web application, which detects the emotion through the web-camera feed and generates a playlist of songs of the preferred genre. Our Convolutional Neural Network (CNN) model is inspired by the Xception architecture combined with the use of residual modules and depth-wise separable convolutions, reducing the number parameters for a faster real-time system. Important concepts such as data augmentation and transfer learning were explored. Our dataset consists of images from the available datasets like CK+, JAFFE, KDEF, FER2013 and a few manually captured photos. Our results show overall 87% accuracy.

**Keywords** — *machine learning, artificial neural networks, deep learning, convolutional neural networks, computer vision, data augmentation, transfer learning.*

## **LIST OF FIGURES**

Figure 1: System Architecture .....	19
Figure 2: DFD 0.....	21
Figure 3: DFD 1.....	22
Figure 4: Use Case.....	23
Figure 5: Class Diagram .....	25
Figure 6: Sequence Diagram.....	26
Figure 7: Image Preprocessing 1.....	28
Figure 8: Image Preprocessing 2.....	28
Figure 9: Image Preprocessing 3.....	29
Figure 10: Transfer Learning.....	30
Figure 11: CNN 1.....	33
Figure 12: CNN 2.....	35
Figure 13: MiniXception Architecture.....	36
Figure 14: Sample Outputs.....	38
Figure 15: Accuracy Comparison .....	41
Figure 16: Screenshots of Web Application.....	47

## **TABLE OF CONTENTS**

<b>1. Introduction.....</b>	<b>7</b>
1.1. Problem Statement.....	7
1.2. Background.....	7
1.3. Objectives.....	7
1.4. Relevance.....	8
1.5. Project Undertaken.....	8
1.6. Organization of report.....	8
<b>2. Background.....</b>	<b>10</b>
2.1. Introduction.....	10
2.2. Facial Expression Recognition.....	10
2.3. Application of AFER in music.....	11
2.4. Past Work.....	11
2.4.1. Feature Extractors.....	12
2.4.2. Classifiers and Predictions.....	12
2.4.3. Facial Emotion Recognition.....	12
2.5. Literature Survey.....	13
<b>3. Requirement Specification.....</b>	<b>16</b>
3.1. Problem Description.....	16
3.2. Scope.....	16
3.3. Assumptions.....	17
3.4. Functional Requirements.....	17
3.5. Non Functional Requirements.....	18
3.6. Hardware and Software Requirements.....	18
<b>4. Design and Analysis.....</b>	<b>19</b>
4.1. System Architecture.....	19
4.2. Data Flow Diagram.....	21
4.2.1. DFD0.....	21
4.2.2. DFD1.....	22
4.3. Use Case Diagram.....	23
4.4. Class Diagram.....	25
4.5. Sequence Diagram.....	26
<b>5. Implementation.....</b>	<b>27</b>
5.1. Introduction.....	27
5.2. Face Detection.....	27

5.3.	Image Preprocessing.....	27
5.4.	Feature Extraction using Local Binary Pattern.....	29
5.4.1.	Drawback of feature extraction.....	30
5.5.	Transfer Learning.....	30
5.5.1.	Pre-trained Model.....	31
5.5.2.	Why use pre-trained model.....	31
5.5.3.	Fine tuning.....	31
5.5.4.	Bottleneck.....	32
5.6.	CNN.....	33
5.6.1.	MiniXception.....	36
<b>6.</b>	<b>Result and Evaluation.....</b>	<b>39</b>
6.1.	Testing and Evaluation.....	39
6.2.	Test Driven Development.....	39
6.3.	Comparative Study.....	40
6.4.	Screenshots.....	43
6.4.1.	Home page.....	43
6.4.2.	Register.....	44
6.4.3.	Login.....	45
6.4.4.	Face Capture.....	46
6.4.5.	Music Player.....	47
<b>7.</b>	<b>Conclusion and Future Work.....</b>	<b>48</b>
7.1.	Conclusion.....	48
7.2.	Future Scope.....	49
	<b>PLAGIARISM REPORT.....</b>	<b>50</b>
	<b>REFERENCES.....</b>	<b>51</b>

## 1. INTRODUCTION

## **1.1 PROBLEM STATEMENT**

Implementing an emotion based music player which detects the current mood of the user utilizing the web-camera video feed and plays a playlist of songs of the preferred genre by the user.

## **1.2 BACKGROUND**

Music plays an important role in everyone's life. Recent studies show listening to music has a certain impact on the activities of human brain. In today's world, with ever increasing advancements in the field of multimedia and technology, music listening experience can be enhanced using machine learning. Various music players are available in the market with features like fast-forward, reverse, variable playback speed, artist classification, genre classification, shuffle etc. Although these features satisfy the user's basic requirements, but still the user must manually browse through the playlist of songs and select songs based on his current mood.

## **1.3 OBJECTIVES**

By undertaking this project we look to provide a new and advanced music listening experience that will also reduce the manual work of scrolling through a playlist and selecting a song. Each user has different preferences of music genre for their different emotional states, availing such a music player application will not only save time but also provide a better human to machine interactions.

## **1.4 RELEVANCE**

Artificial Intelligence is the next big Industrial Revolution. It has led to the creation of not only smart devices but also smart environments. Thus, there is also a need for a smart and intelligent music player. It should allow its users to set



preferences and provide a better experience by automating the manual interactions. The emotion based music player will provide a new platform to all music listeners.

## **1.5 PROJECT UNDERTAKEN**

Emotion based music player project lies under the domain of artificial intelligence. The aim of the project is to determine human emotion using facial expressions and playing suitable playlist of songs.

Deep Neural Networks are one of the most powerful tools of Artificial Intelligence, with excellent performances in complex machine learning problems. Especially the applications dealing with image data, such as largest image classification data set (Image Net), computer vision, and in natural language processing (NLP), surpassing the traditional Artificial Neural Networks. So hence, we propose a deep learning approach to solve the problem. Deep neural networks such as CNNs have proved effective in solving many complex machine learning problems especially in the field of computer vision under artificial intelligence.

The project is split into two halves; using the deep learning model to determine the emotion of the user and generating a playlist of the preferred genre of songs. Considering the practicality of the project, we have considered 4 different emotions i.e happiness, anger, sadness and neutral.

## **1.6 ORGANIZATION OF THE REPORT**

This project report contains 7 chapters. The first chapter gives an introduction of the project including the problem statement, relevance and past related work. The second chapter gives the background which includes previous similar systems and literature survey. Third chapter is specifications which gives information about the hardware and software requirements. The fourth chapter is design which tells about the overall design of the modules and architecture of the project. Fifth chapter explains the implementation of the project. Sixth chapter gives you the results which we have achieved and contains evaluation of the same. The last chapter gives a conclusion and future scope of our project.

## **2. BACKGROUND**

### **2.1 INTRODUCTION**

This chapter provides a brief overview of the automatic facial expression recognition system and previous studies done in the field of facial expression recognition, giving an overview of various tools and technologies used earlier. It explains the applications of facial expression recognition in the field of music and musical technologies. Literature survey of the various sources used in the development of the project are also listed.

## **2.2 FACIAL EXPRESSION RECOGNITION**

Facial expressions is one of the primary mediums for expressing feelings and emotions. Universally, there are seven facial expressions happy, anger, contempt, disgust, fear, sadness and surprise. In communication, along with voice, facial expressions play a vital role in transmitting information about a person's emotional, mental as well as physical state. Thus, facial expressions have been studied by researchers for a long time. Different fields like psychology, behavioral science and bio-metric area of research study and use facial expressions to understand human psychology.

This gives a rise to study of facial expressions in the field of computer science. Technologies such as artificial intelligence, machine learning and computer vision has allowed researchers to develop systems that can automatically recognize facial expressions and predict the emotion. Various applications like emotion detection, human computer interaction, lie detection, mental state tracking, intelligent teaching system etc. can use automatic facial expression recognition system (AFERS).

Developing an AFER that is robust, person independent and that ideally works in real time scenarios requires various constraints like image quality, illumination of the face, distance of camera from face etc. have to be considered in order to achieve good results. Hence, machine learning algorithms and computer vision technologies play an important role to achieve the best possible results.

## **2.3 APPLICATION OF AFER IN MUSIC**

Music is a personal choice and plays a prominent role for each individual. It is one of the basic forms of acoustic entertainment and can be easily accessed by

everyone. Most of the music listeners tend to have a large playlist of songs in their music player. Generally people have a habit of listening to music according to their current emotion and mood. For this, they have to manually go through their long playlist and find a suitable song as per their emotion.

Advancements in technology such as Artificial Intelligence, automating such manual processes has gained a lot of importance to improve the user experience as well as save time. A smart music player that will recognize human emotion and automatically play suitable songs as per the preference will provide a more implicit human computer interaction.

Emotion based music player is an application that will predict current emotion of the user utilizing the web-camera video feed. After detecting the emotion, it will play a playlist of songs of the preferred genre of choice of the user.

The emotion based music player will be advantageous to users looking for music based on their mood and emotional behaviour providing a better user experience. The system can also be helpful in music therapy treatment and provide the music therapist the assistance needed to treat the patients suffering from disorders like mental stress, anxiety, acute depression and trauma. It also has many other abundant potential applications such as human computer interaction (HCI), behavioral science, psychology, video games etc.

## 2.4 PAST WORK

Facial Action Coding System (FACS) is a taxonomy of human facial expressions, and is the most commonly used system to objectively describe the facial expression signal by human observers. It currently specifies 32 atomic facial muscle actions, named Action Units (AU) and was the earliest approach to solve the problem.

The Computer Vision community has defined a set of problems related to the automatic analysis of AU, such as AU detection, AU intensity estimation, and the automatic detection of the AU temporal segments. It's very time consuming, and in addition annotators require expert training to be able to produce consistent annotations. Furthermore, the annotation of AU intensities is particularly challenging given the small variation between consecutive levels.

Another approach is the dimensional approach which utilizes geometric and appearance features to predict the expression. Geometric features encode

information based only on the facial landmark locations, appearance features encode pixel intensity information instead, and motion features are constructed based on a dense registration of appearances between (consecutive) frames. However this approach requires accurate and reliable facial landmarks detection, along with tracking methods and often sensitive to noise.

The generalized approach in solving the problem of emotion recognition has three major parts. To extract and determine the emotion of a user effectively, we need to extract features from an image and use them against a trained data set to classify the input and determine the emotion.

### **2.4.1 Feature Extractors:**

A feature extractor is an application which extracts important points in an image and generates a feature vector. For emotion recognition facial regions such as eyes, nose, lips and chin are segmented, its feature vector such as local binary pattern (LBP) is computed and concatenated together.

### **2.4.2 Classifiers and Prediction:**

After extracting features from an image set of training and testing data, a feature classifier is needed to sort out and classify the testing data with relevance to the training data.

### **2.4.3 Facial Emotion Recognition:**

Several approaches have been proposed to classify human affective states. The features used are typically based on displacements of specific points or spatial locations of particular points; this technique is known as Facial Action Coding System (FACS). Later, several machine learning models trained with the extracted facial features gave better results but still required considerable computation time.

Our approach is using deep learning, deep neural networks have the ability to learn the required features automatically but generally requires a larger dataset for training. We created a dataset of images by combining images from various available datasets for facial expressions and few pictures captured by a web

camera. In order to have enough images for a deep neural network we needed to apply the concept of data augmentation which allows to perform transformations on the available images to create a larger dataset. Furthermore, the concept of transfer learning was explored which allows to create better models in the case of smaller size of dataset. It utilizes a pre-trained model and further learns to perform a new task by training on the smaller dataset. Finally, we created a CNN inspired by Xception architecture combined with the use of residual modules and depth-wise separable convolutions, resulting into the reduction of number parameters for a faster real-time system.

## 2.5 LITERATURE SURVEY

The work of Dr. Valstar and Dr. Martinez is funded by European Union Horizon 2020 research and innovation programme gives a complete review of Automatic Facial Expression System. This paper discusses the various methods, challenges and opportunities in AFER.[1]

Anima Majumder et al proposed a feature fusion framework using autoencoders which can learn the correlation and generate a better representation of both, geometric and appearance features. Viola Jones is used for face detection, geometric features were extracted and encoded with LBP based appearance features to form a combined representation. Fused feature is applied to SOM based classifier and performance was validated on two widely used databases (DBs): 1) MMI and 2) extended CohnKanade (CK+).[2]

As there is excellent capability of description of local texture, local binary patterns (LBP) have been applied in many areas. In this paper, they have enhance the classical LBP method from three aspects for facial expression recognition: image data, extracting features and the way of combining all these features. At last, we adopt wavelet to decomposed image. Then they extracted LBP features with a new local and holistic way to make features more robust. At last, in order to use the extracted features more logical. Proposed improvements in this paper have promoted the performance of facial expression recognition greatly.[3]

The study of music and emotions suggests that there is a psychological relationship between a person's emotional state and the type of music they listen to. Use of machine learning and data mining is widely done in making emotion based music player. extraction of genre is important to play the right song according to user preference.[4]

Deep learning networks have been utilized in solving complex machine learning problems. CNNs are one of the most widely used deep neural network. This paper establishes the required background and concepts of CNN.[5]

Cahit Deniz GÜRKAYNAK et al performed a case study on transfer learning approach in convolutional neural networks. Transfer learning parameters are examined on AlexNet, VGGNet and ResNet. Their results confirmed that transferring the parameter values of the first layers and fine-tuning the other layers, whose weights are initialized from pre-trained weights, performs better than training network from scratch. It's also observed that preprocessing and regularization improves overall scores significantly.[6]

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun et al present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. They have explicitly reformulated the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. They provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset they have evaluated residual nets with a depth of up to 152 layers. [7]

Depthwise separable convolutions reduce the number of parameters and computation used in convolutional operations while increasing representational efficiency. They have been shown to be successful in image classification models. both in obtaining better models than previously possible for a given parameter count and considerably reducing the number of parameters required to perform at a given level (the MobileNets family of architectures). Depthwise separable convolutions can be applied to neural machine translation. [8]

A properly trained system is necessary for facial expression recognition. For getting good accurate results, preprocessing of the image has to be done properly. Illumination of the image has been a basic problem for any image based application. Incorrect illumination causes the system to interpret the image in a wrong manner. Using various methods to improve the illumination and the quality of the image helps to achieve better results. [11]

### **3. REQUIREMENT SPECIFICATION**



### 3.1 PROBLEM DESCRIPTION

Implementation of an emotion based music player which will detect emotion from a live video feed of a user and play songs according to the detected emotion through the following steps :-

1. From the webcam feed of user's device, frames will be captured.
2. User's face is detected in the frame.
3. Facial region image is extracted from the frame.
4. Various pre-processing techniques are applied on that image to improve its quality.
5. The image acts as an input to a neural network which determines the emotion.
6. The detected emotion will be mapped to the music genre preference of the user.
7. Playlist of the mapped genre is returned and the songs are played.

### 3.2 SCOPE

Facial expressions have always been a great indicator of the state of mind of a person. The most natural way to express emotions is through facial expressions. Facial expression recognition has played a great role in fields of HCI to determine the behavior of a user with respect to the software, in gaming to determine how the gamer's expression changes during various scenarios of the game, in security to determine whether a person possesses a threat or not.

The emotion based music player will be advantageous to users looking for music based on their mood and emotional behaviour providing a better user experience. The system can also be helpful in music therapy treatment and provide the music therapist the assistance needed to treat the patients suffering from disorders like mental stress, anxiety, acute depression and trauma. The emotion based player will be of great advantage to users :

1. Looking for music, based on their mood.
2. Wanting a music player which behaves according to emotion.

### 3.3 ASSUMPTIONS

1. The user must be in a good lighting condition.
2. The camera resolution should be atleast 2MP.

### 3.4 FUNCTIONAL REQUIREMENTS

The following are the functionalities in our system:

#### 1. Face Detection -

First a frame is captured through the live feed of the webcam of the device. The image will be passed on to a face detector module which give the image of the face and discard the remaining image.

#### 2. Preprocessing -

This module preprocesses the image by performing operations to improve the contrast of the image, aligning the face correctly, denoising the image etc. so the the image will be ready to be passed to the neural network.

#### 3. Emotion Recognition -

This module uses the preprocessed image to determine the emotion using a trained neural network.

#### 4. Music Player -

This module plays a song by mapping the detected emotion to the genre of the song. Users have to provide their genre preferences during the registration process ie. they will specify the genre which they like to listen to when they are in a particular mood. The emotion is mapped to their preference returning a playlist.

### 3.5 NON-FUNCTIONAL REQUIREMENTS

**Availability** - Our software will be on a server; hence it is available anytime.

**Usability** - A user can access the software through any device and a browser.

### **3.6 HARDWARE AND SOFTWARE REQUIREMENTS**

- **Hardware requirements :**

4GB RAM

Camera 2MP

2GB Graphic Card

Intel i3 (4th Gen)

- **Software requirements :**

Python 3.x and required libraries

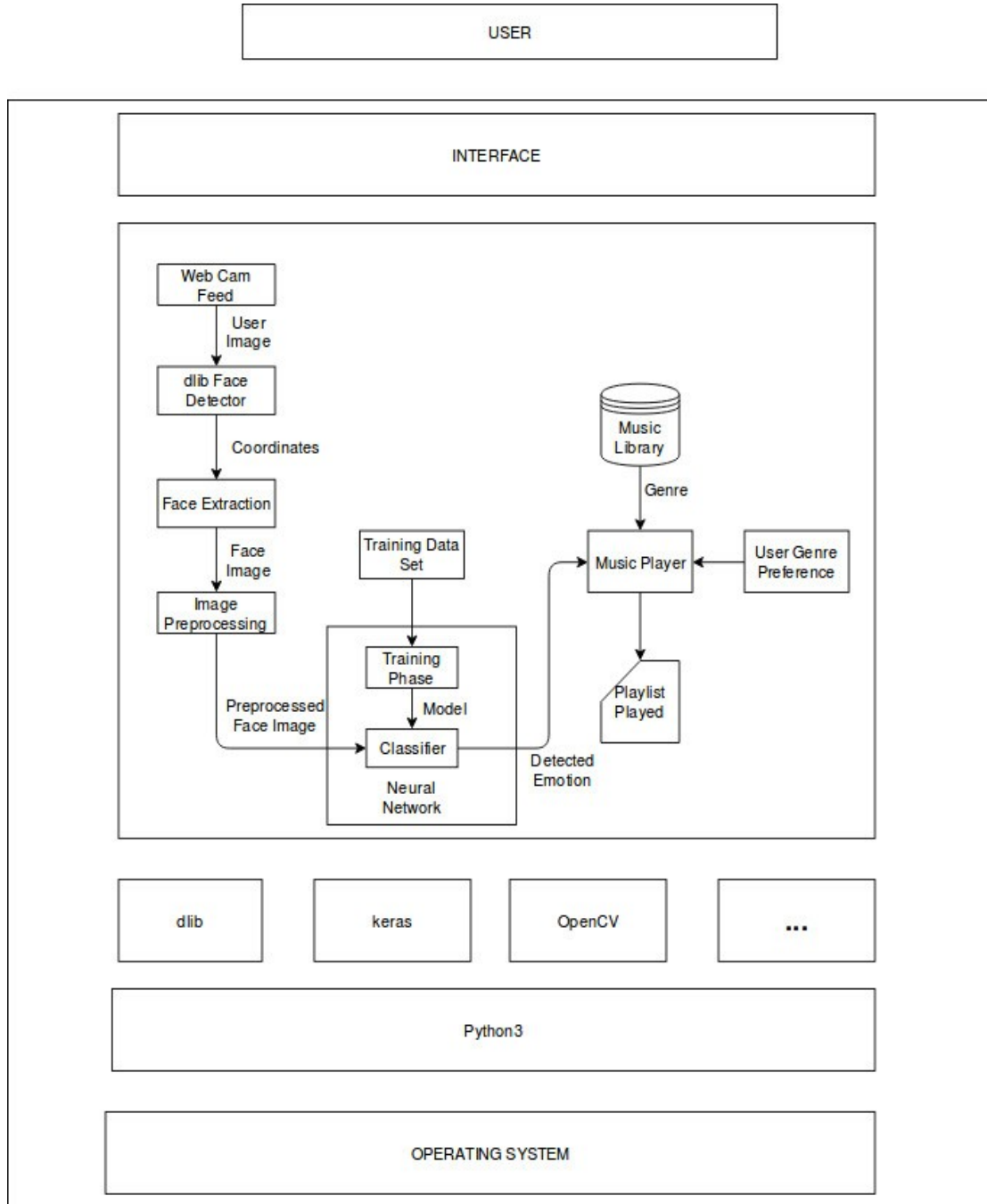
Linux/Windows OS

Chrome/Firefox browsers

MySQL database

## **4. DESIGN AND ANALYSIS**

## 4.1 SYSTEM ARCHITECTURE



( Fig 1 : SYSTEM ARCHITECTURE )

The system architecture diagram shows the overall modules and working of the project. Following is the explanation of the core working section our architecture.

- First, a frame of the user is captured by a live feed of web camera and is given as an input to the pre trained dlib face detector module. The module gives coordinates of the face from the frame.
- Next, the face extraction module uses these coordinates to extract the face from the frame and discard the remaining image.
- This extracted face is passed to the image preprocessing module which makes the image ready for emotion detection.
- The face is passed on to the classifier which will determine the human emotion. This completes the human emotion detection phase.
- Next, there is a music player which maps the detected emotion to the user's preference of music genre i.e. which is the genre preference of the user when he is happy or sad or angry etc.
- Finally, the music player returns a playlist of that genre and plays a song.

The entire system is implemented using the python programming language along with the help of its various packages and libraries.

At the top, there is an interface, which is built using HTML5, CSS, bootstrap, and at the end flask web application connects our system with the user.

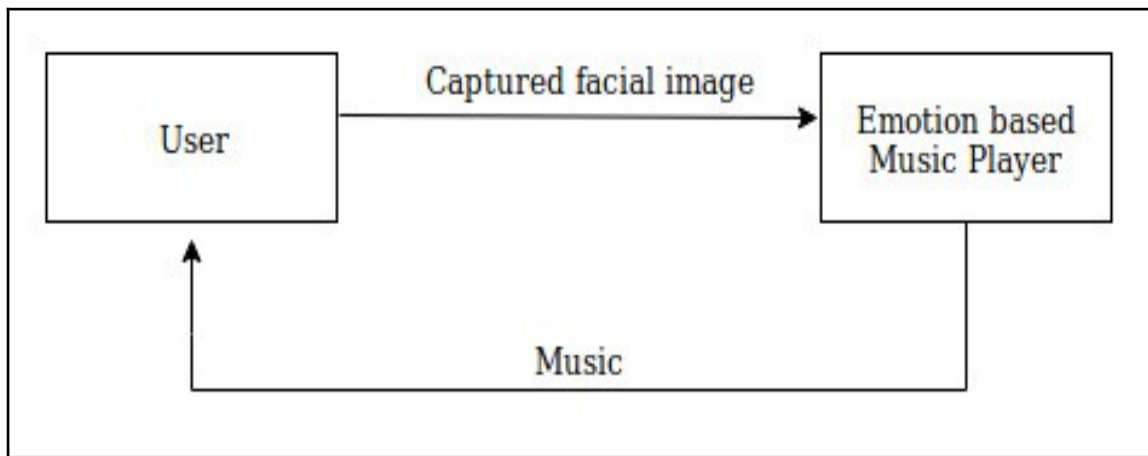
## 4.2 DATA FLOW DIAGRAM (DFD)

A **data flow diagram (DFD)** is a graphical representation of the "flow" of data through an information system, modelling its *process* aspects. A DFD shows what kind of information will be input to and output from the system, how the data will advance through the system, and where the data will be stored. This diagram shows the flow starting from the image being captured to the song being played.

### 4.2.1 DFD 0

DFD 0 gives a very basic idea of how the data flows in our system. User sends a frame from live feed, the player returns a song based on the emotion.

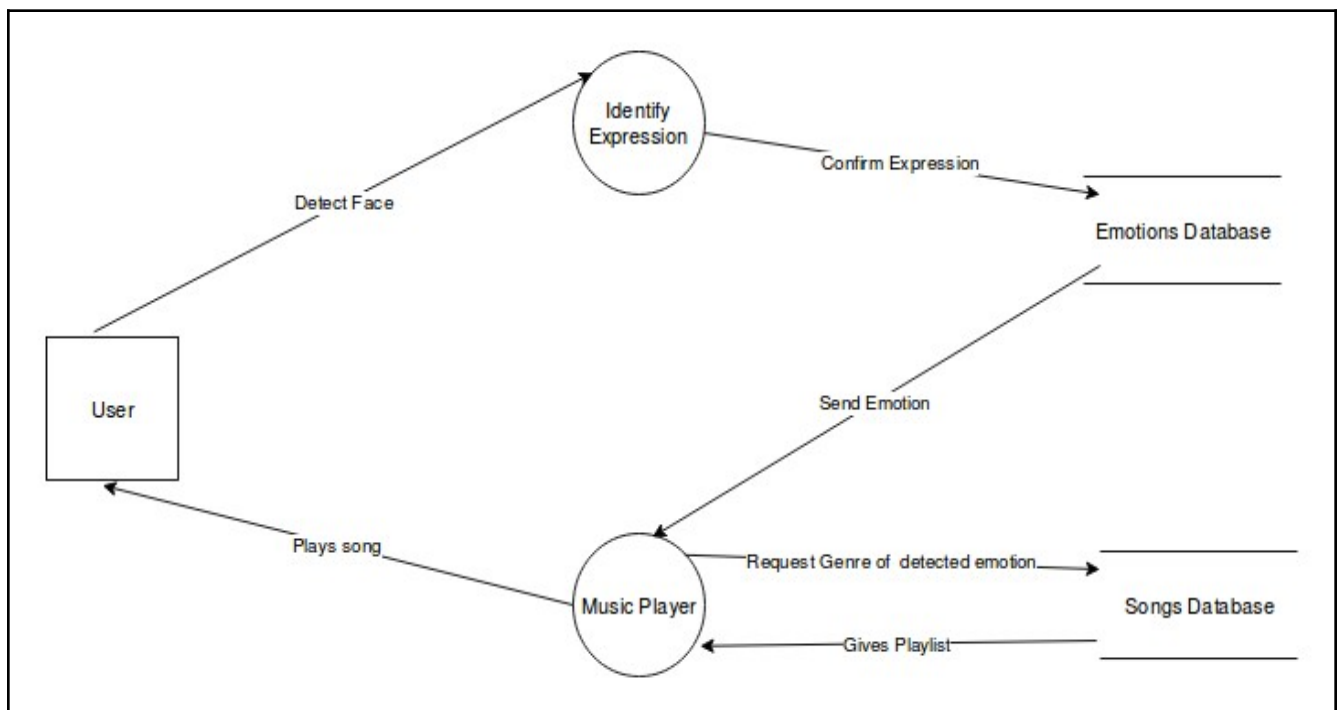
## Emotion Based Music Player



( Fig 2 : DFD 0)

### 4.2.2 DFD 1

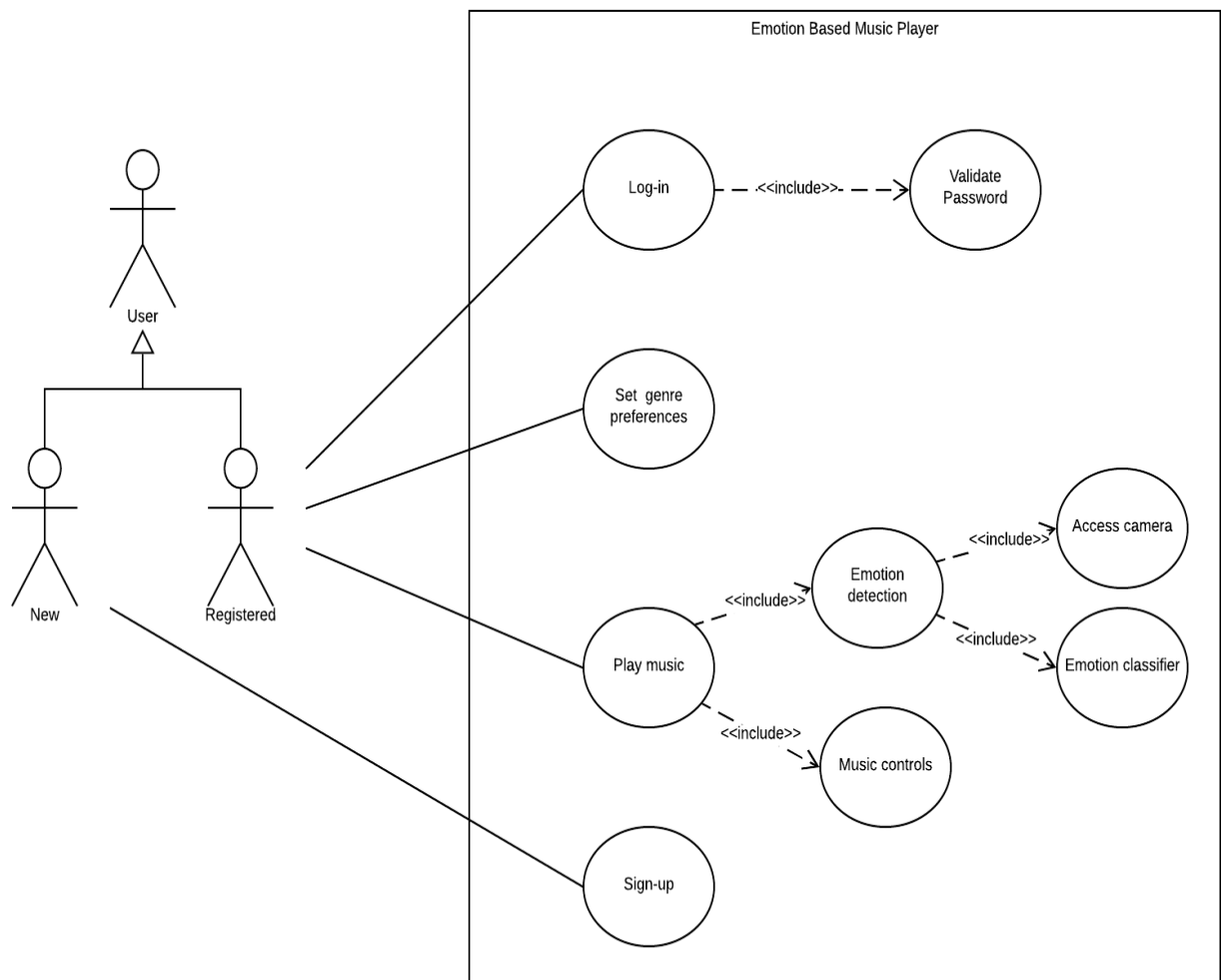
DFD 1 shows the details of the data being passed on from one module to another.



( Fig 3 : DFD 1)

### 4.3 USE CASE DIAGRAM

Use case diagrams are used to specify external **requirements**, required usages of a system under design or analysis (**subject**) - to capture what the system is supposed to do; the **functionality** offered by a subject – what the system can do; requirements the specified subject poses on its **environment** - by defining how environment should interact with the subject so that it will be able to perform its services.



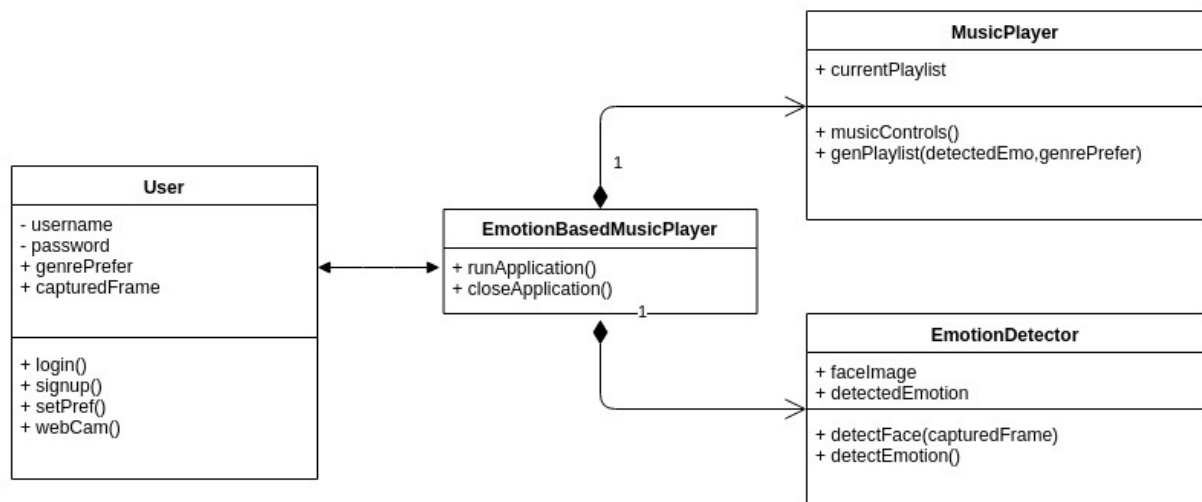
( Fig 4 : USE CASE DIAGRAM )

This is the use case diagram of our system. A user can be either a new user or an existing user. If the user is a new user, he is to sign up first. An existing user has to login to his account using his credentials. Once the user decides to play music, the system captures an image, detects the emotion and plays the song. The system gives a playlist and user has the ability to control the music.



## 4.4 CLASS DIAGRAM

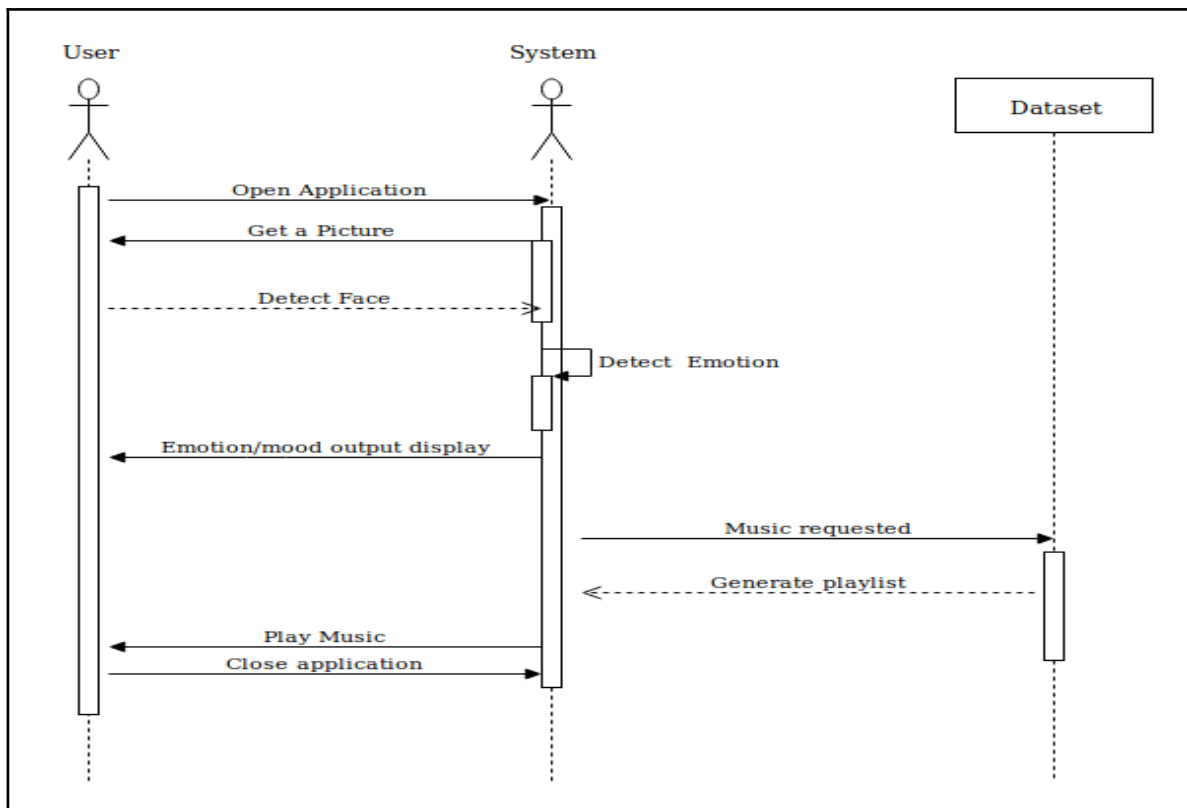
Class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects. class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.



( Fig 5 : CLASS DIAGRAM )

## 4.5 SEQUENCE DIAGRAM

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. A sequence diagram shows, as parallel vertical lines (*lifelines*), different processes or objects that live simultaneously, and, as horizontal arrows, the messages exchanged between them, in the order in which they occur. This allows the specification of simple runtime scenarios in a graphical manner.



(Fig 6 : SEQUENCE DIAGRAM)

This is the sequence diagram of our system. The user opens the application. The system will click the picture of the user through a live feed of webcam. It will detect the emotion and will ask the dataset to return a playlist according to the detected emotion. A song will be played and the controls will be handed to the user.

## **5. IMPLEMENTATION**

### **5.1 Introduction**

This chapter provides a brief overview of the various techniques used for automatic facial expression recognition, their implementation and drawbacks if any. It begins with the mandatory modules, which are used in all methods for automatic facial expression recognition then dealing with different modules of various techniques.

### **5.2 Face Detection**

First step in emotion detection is to extract the useful section of face of a human being so that it is easier to work for classification process.

Face is detected from continuous frames received through webcam using video capture class of openCV. The class provides C++ API for capturing video from cameras.

Face is then detected using dlib's frontal face detector.

### **5.3 Image Preprocessing**

It includes performing of various operations on an image to improve image data so that the image is enhanced and the unwanted data or the noise in the image is removed.

**Various Techniques :**

1. **Face Alignment** : Image is rotated that such the eyes lie on a horizontal line that is along the same y-coordinates.



( Fig 7 : PREPROCESSING 1 )

2. **Image Denoising** : It is the process of removing unwanted noise from the image to restore a proper image.



( Fig 8 : PREPROCESSING 2 )

- 3. Contrast Improvement :** Contrast Limited Adaptive Histogram Equalization (CLAHE) is used to improve the contrast of the image.



( Fig 9 : PREPROCESSING 3)

- 4. Resize Image :** Some images captured by a camera and fed to AI algorithm vary in size, therefore, we should establish a base size for all images fed into AI algorithms.

Above two modules are necessary for all methods used for automatic facial expression detection. Now we will discuss the various methods one by one which are used for facial expression detection.

## 5.4 Feature Extraction using Local Binary Patterns

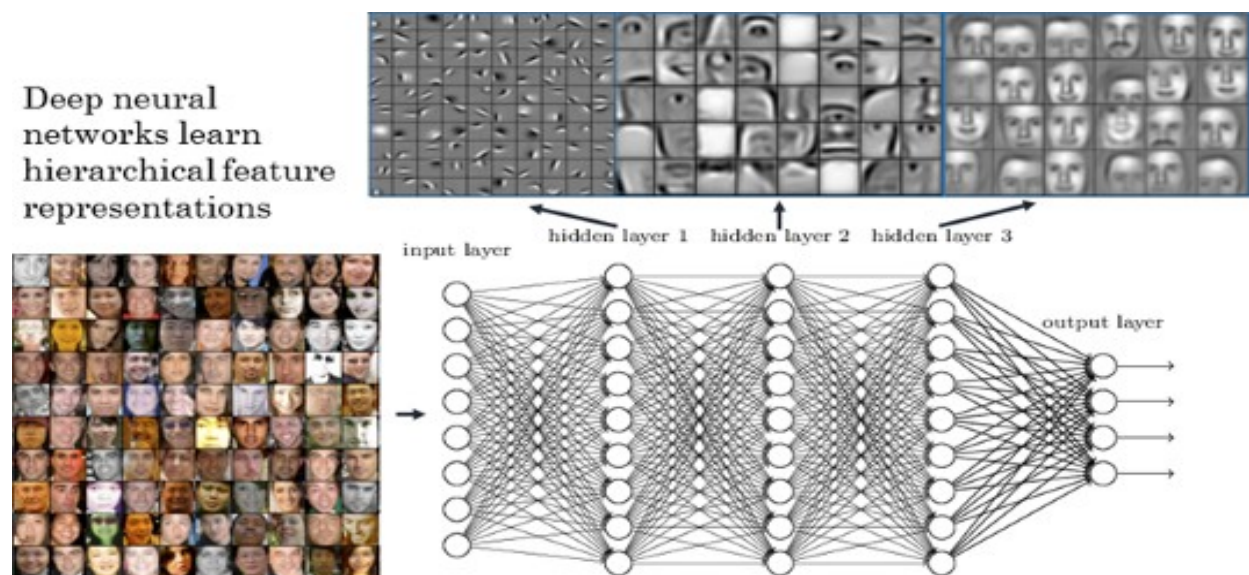
LBP is an efficient operator for texture representation of Image pattern. LBP involves dividing up the face image to regions. The neighbouring pixels are examined based on the central pixel grayscale value which threshold the neighbours to 1 or 0. Hence a binary string representing each pixel will be formed. Histograms of 256 bins for every region will be build which are concatenated to form the feature vector for the face image

Neural network is applied as a classifier in this system. Input to the training phase is collection of images having feature vectors of human faces. Output is a score against a class which indicates the emotion which is detected by the model.

### 5.4.1 Drawbacks of feature extraction

1. Feature extraction cannot be tweaked according to the classes and images.
2. Feature extraction is done in an unsupervised manner where in the classes of the image have nothing to do with information extracted from pixels.
3. If the chosen feature lacks the representation required to distinguish the categories, the accuracy of the classification model suffers a lot, irrespective of the type of classification strategy employed.

### 5.5 Transfer Learning



( Fig 10 : TRANSFER LEARNING )

Transfer learning involves the approach in which knowledge learned in one or more source tasks is transferred and used to improve the learning of a related target task. In transfer learning we first train a base network on a base dataset and task, and then we repurpose the learned features, or transfer them, to a second target network to be trained on a target dataset and task.

### 5.5.1 Pre-trained Model

A pre-trained model has been previously trained on a dataset and contains the weights and biases that represent the features of whichever dataset it was trained on. Learned features are often transferable to different data. For example, a model trained on a large dataset of bird images will contain learned features like edges or horizontal lines that would be transferable to your dataset.

### 5.5.2 Why use a Pre-trained Model

Pre-trained models are beneficial to us for many reasons. By using a pre-trained model you are saving time. Someone else has already spent the time and compute resources to learn a lot of features and your model will likely benefit from it.

Pre-trained models can be used when your dataset is less than required for a training a CNN from scratch.

### 5.5.3 Fine Tuning

Since the dataset is small, we have done data augmentation to increase dataset size and we try to fine-tune through the full network.

#### **Data Augmentation :**

It is done via a number of random transformations, so that our model would never see twice the exact same picture. This helps prevent overfitting and helps the model generalize better. The transformations include randomly rotate pictures,

randomly translate pictures vertically or horizontally, applying shearing transformations and randomly zooming inside pictures

### **Fine Tuning is performed as follows :**

1. Remove the last fully connected layer and replace with the layer matching the number of classes in the *target* dataset
2. Randomly initialize the weights in the new fully connected layer
3. Initialize the rest of the weights using the pre-trained weights, i.e., unfreeze the layers of the pre-trained network
4. Retrain the entire neural network

### **Fine tuning result :**

It did not give good results as Imagenet and Facial Expressions Recognition have very different datasets.

### **5.5.4 Bottleneck**

The basic technique to get transfer learning working is to get a pre-trained model (with the weights loaded) and remove final fully-connected layers from that model. We then use the remaining portion of the model as a feature extractor for our smaller dataset. These extracted features are called "Bottleneck Features" (i.e. the last activation maps before the fully-connected layers in the original model). We then train a small fully-connected network on those extracted bottleneck features in order to get the classes we need as outputs for our problem.

### **Bottleneck Result :**

The validation accuracy and precision were low and validation loss was high, therefore we discarded this method.

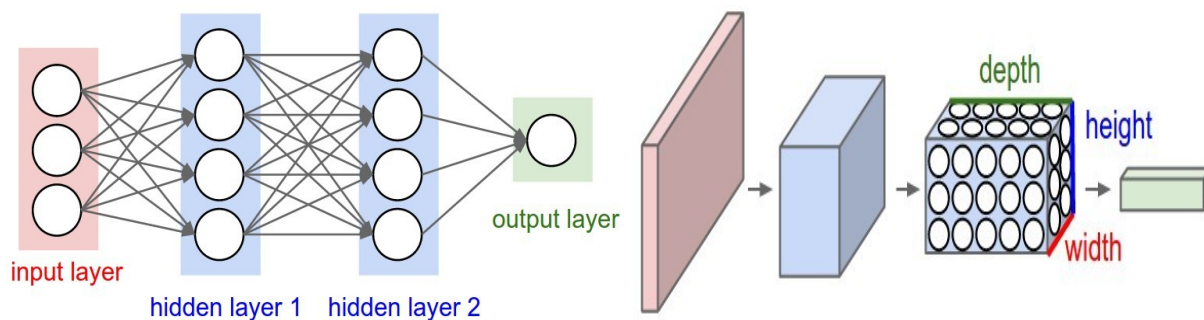


## 5.6 CNN

A Convolutional Neural Network (CNN, or ConvNet) are a special kind of multi-layer neural networks, designed to recognize visual patterns directly from pixel images with minimal preprocessing. With Convolutional Neural Networks (ConvNets), the task of training the whole network from the scratch can be carried out using a large dataset like ImageNet.

The ImageNet project is a large visual database designed for use in visual object recognition software research and currently has 14,197,122 images from 21841 different categories.

Convolutional Neural Networks take advantage of the fact that the input consists of images and they constrain the architecture in a more sensible way. In particular, unlike a regular Neural Network, the layers of a ConvNet have neurons arranged in 3 dimensions: width, height, depth.



( Fig 11 : CNN 1)

A simple ConvNet is a sequence of layers, and every layer of a ConvNet transforms one volume of activations to another through a differentiable function. We use three main types of layers to build ConvNet architectures: Convolutional Layer, Pooling Layer, and Fully-Connected Layer (exactly as seen in regular Neural Networks). We will stack these layers to form a full ConvNet architecture.

## Convolutional Layer

The Conv layer is the core building block of a Convolutional Network that does most of the computational heavy lifting. Convolution is the first layer to extract features from an input image. Convolution preserves the relationship between pixels by learning image features using small squares of input data. Three hyperparameters control the size of the output volume: the depth, stride and padding.

1. Depth : it corresponds to the number of filters we would like to use.
2. Stride : Stride is the number of pixels shifts over the input matrix. When the stride is 1 then we move the filters to 1 pixel at a time. When the stride is 2 then we move the filters to 2 pixels at a time and so on.
3. Padding : Sometimes filter does not fit perfectly fit the input image. We have two options:
  - a. Pad the picture with zeros (zero-padding) so that it fits
  - b. Drop the part of the image where the filter did not fit. This is called valid padding which keeps only valid part of the image.

## Depthwise Separable Convolutions

Depthwise separable convolutions work with kernels that cannot be “factored” into two smaller kernels.

The depthwise separable convolution is so named because it deals not just with the spatial dimensions, but with the depth dimension—the number of channels—as well. This will perform a spatial convolution while keeping the channels separate and then follow with a depthwise convolution.

An input image may have 3 channels: RGB. After a few convolutions, an image may have multiple channels. You can think of each channel as a particular interpretation of that image; for example, the “red” channel interprets the “redness” of each pixel, the “blue” channel interprets the “blueness” of each pixel, and the “green” channel interprets the “greenness” of each pixel. An image with 64 channels has 64 different interpretations of that image.

This type of CNN is widely used because of the following two reasons –

- They have lesser number of parameters to adjust as compared to the standard CNN's, which reduces overfitting
- They are computationally cheaper because of fewer computations which makes them suitable for mobile vision applications

### **Pooling Layer**

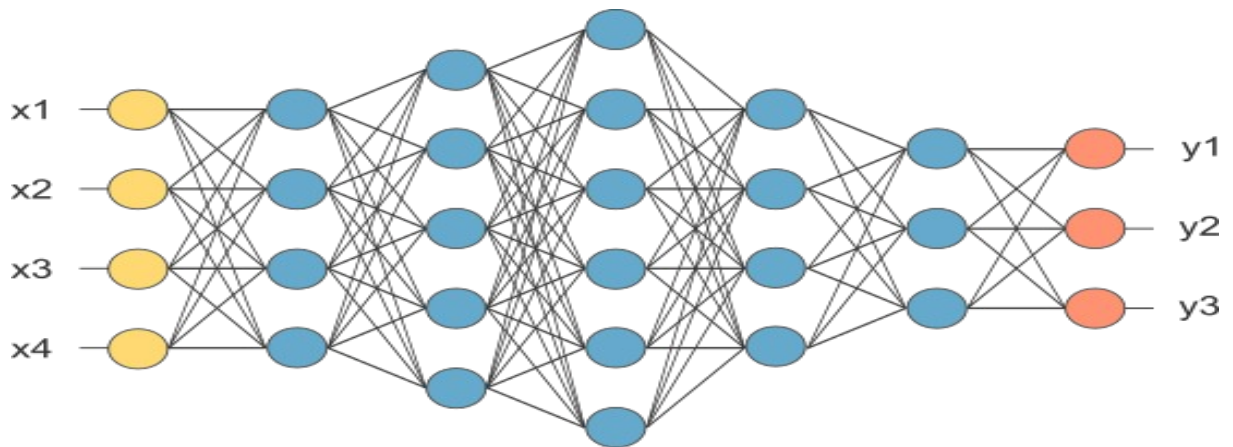
Pooling layers section would reduce the number of parameters when the images are too large. Spatial pooling also called subsampling or downsampling which reduces the dimensionality of each map but retains the important information. Spatial pooling can be of different types:

- Max Pooling
- Average Pooling
- Sum Pooling

### **Fully Connected Layer**

Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in regular Neural Networks. Their activations can hence be computed with a matrix multiplication followed by a bias offset.

The output from the convolutional and pooling layers represent high-level features of the input image. The purpose of the Fully Connected layer is to use these features for classifying the input image into various classes based on the training dataset.

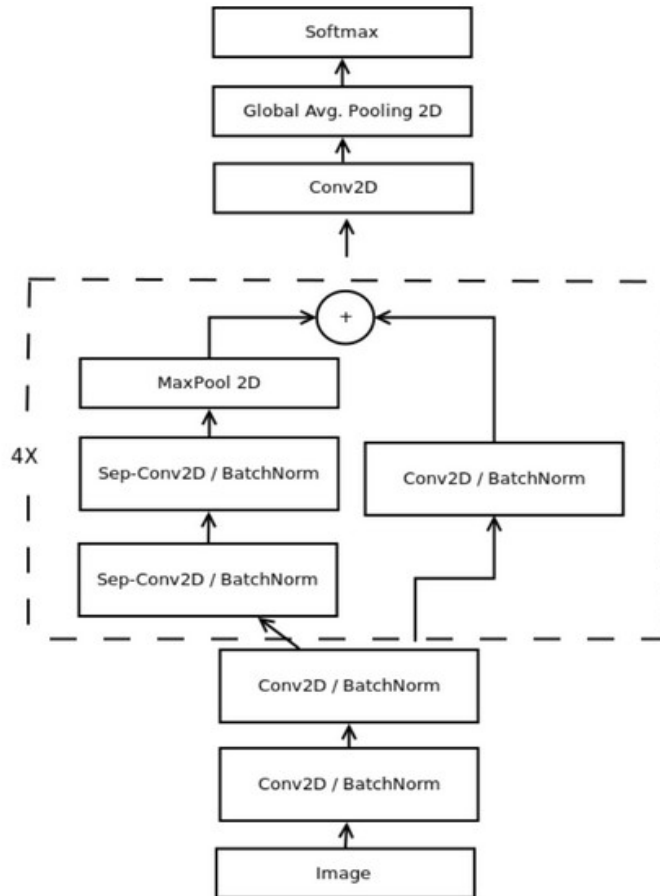


( Fig 12 : CNN 2 )

### 5.6.1 MiniXception

Commonly used CNNs for feature extraction include a set of fully connected layers at the end. Fully connected layers tend to contain most of the parameters in a CNN. Recent architectures such as Inception, reduced the amount of parameters in their last layers by including a Global Average Pooling operation. Global Average Pooling reduces each feature map into a scalar value by taking the average over all elements in the feature map. The average operation forces the network to extract global features from the input image. Modern CNN architectures such as Xception leverage from the combination of two of the most successful experimental assumptions in CNNs: the use of residual modules and depth-wise separable convolutions. Depth-wise separable convolutions reduce further the amount of parameters by separating the processes of feature extraction and combination within a convolutional layer.

#### Architecture



( Fig 13 : MINI XCEPTION ARCHITECTURE )

Xception architecture combines the use of residual modules and depth-wise separable convolutions. Residual modules modify the desired mapping between two subsequent layers, so that the learned features become the difference of the original feature map and the desired features.

The architecture deleted the last fully connected layer, we reduced further the amount of parameters by eliminating them now from the convolutional layers. This was done through the use of depth-wise separable convolutions. Depth-wise separable convolutions are composed of two different layers: depth-wise convolutions and pointwise convolutions. The main purpose of these layers is to separate the spatial cross-correlations from the channel cross correlations.

Depth-wise separable convolutions reduces the computation with respect to the standard convolutions.

It is a fully convolutional neural network that contains 4 residual depth-wise separable convolutions where each convolution is followed by a batch normalization operation and a ReLU activation function. The last layer applies a global average pooling and a softmax activation function to produce a prediction.

There are total 28 layers in our architecture, because the validation accuracy was maximum with 28 layers. It saturated after that and was less with lesser layers.

### Optimization Configuration

Optimizer : adam

Batch Size : 32

Epochs : 200

Initial learning rate: 0.001

### Training

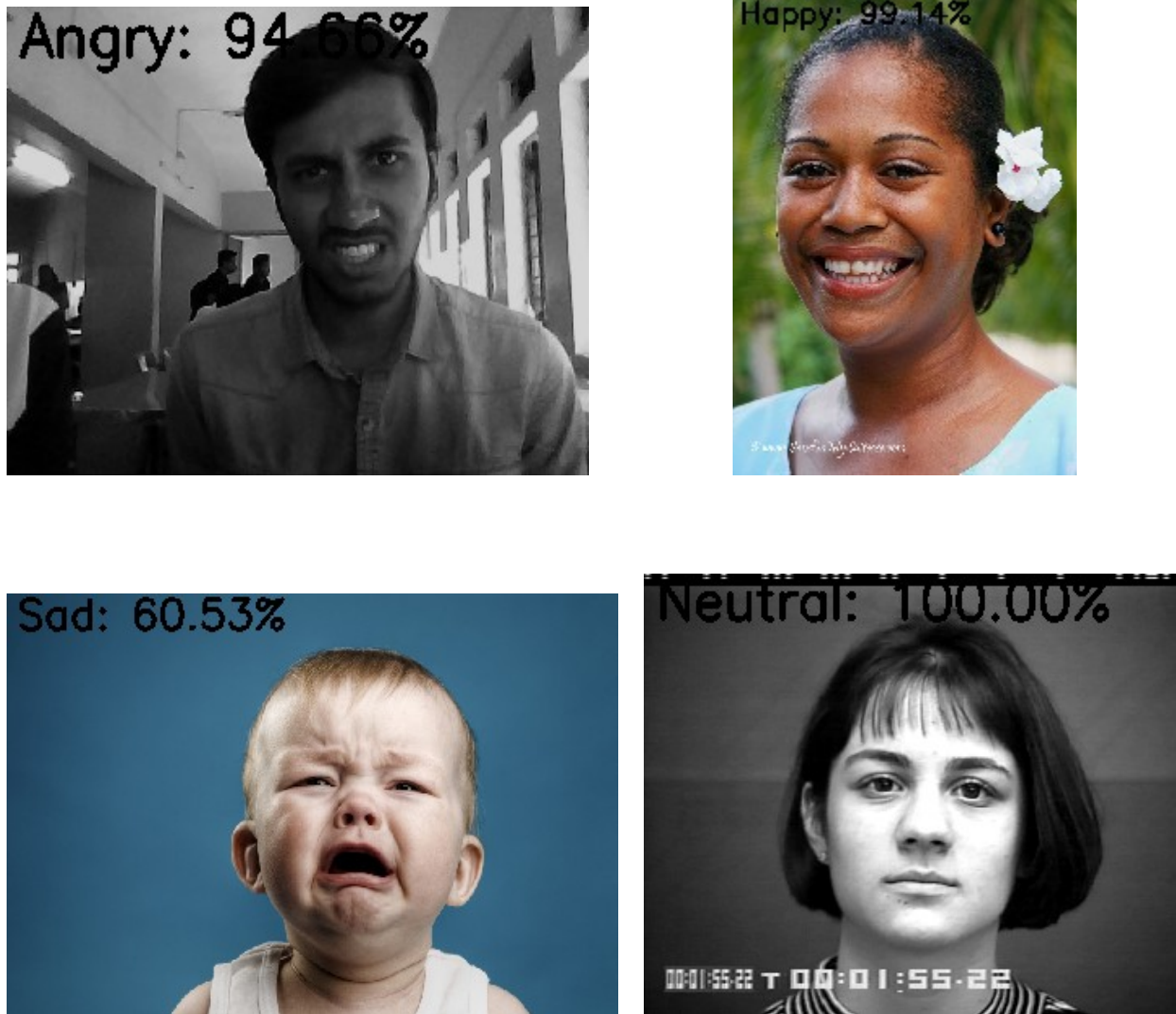
Trained on 14302 samples, validated on 4768 samples

loss: 0.0422 - acc: 0.9984 - val\_loss: 0.1650 - val\_acc: 0.8848

### Output

The MiniXception has given a maximum overall accuracy of 87%. With individual accuracies as follows :

Angry	82%
Happy	100%
Neutral	89%
Sad	78%



( Fig 14 : SOME SAMPLE OUTPUTS)

## 6. RESULT AND EVALUATION

### 6.1 Testing and evaluation

The evaluation of the software indicates that the primary objective of the system has been achieved using multiple techniques to perform emotion

recognition and determining which has the highest success rates. Also required music playlist has been mapped with the resultant emotion of the user. Testing the developed system was done with around 19070 images from various databases. The images were tested using various neural networks. Each technique portrayed different results. The emotions considered were:

1. Angry
2. Happy
3. Neutral
4. Sad

### 6.2 Test Driven Development:

The entire system was designed with a test driven development approach. Final test cases of people portraying particular emotions was determined. A set of 14302 images were used for training and validated on 4768 images. Approximately 4270 images for each emotion were used to train the system thoroughly.

### 6.3 Comparative Study

After training various model architectures and comparing their results, the model with the best outcomes was MiniXception CNN.

**Dataset Used :** KDEF, CK+, JAFFE, FER2013

**Number of samples :** 19070 images

**Number of folders in dataset :** 4

**Categories :** Angry, Happy, Neutral, Sad



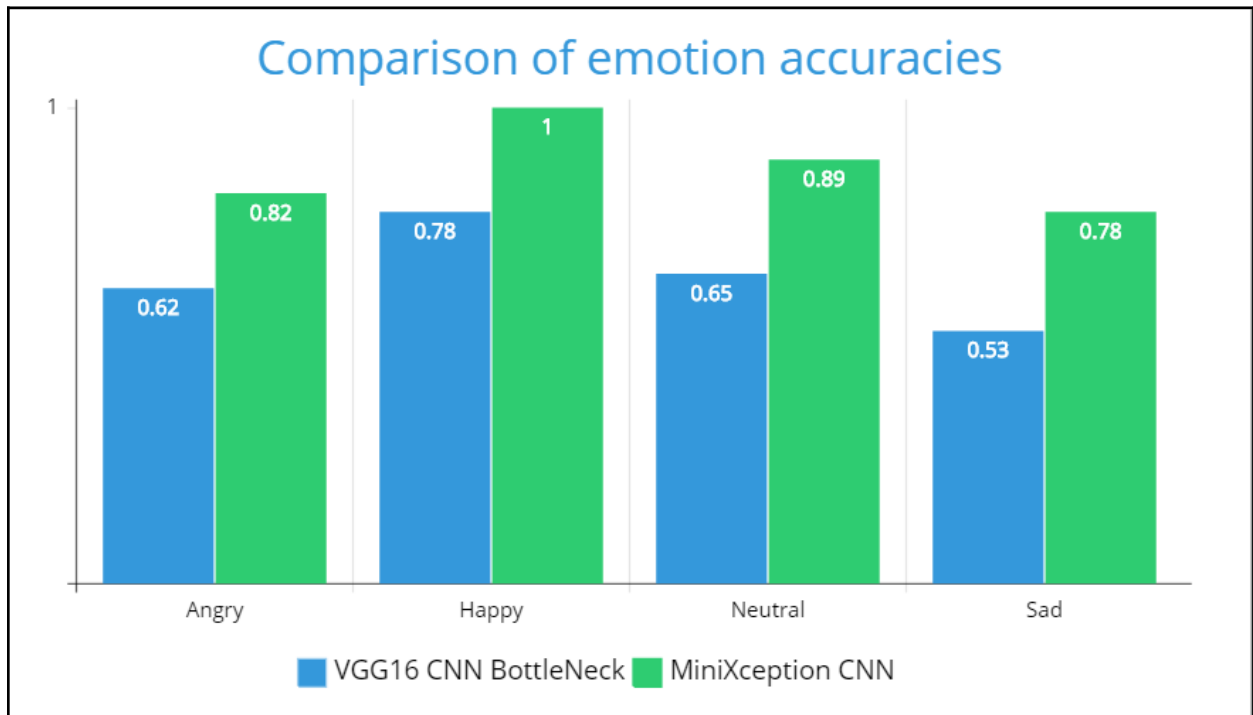
The comparative study of MiniXception CNN with VGG16 CNN BottleNeck is as follows :

	Architecture Specification	VGG16 CNN BottleNeck	MiniXception CNN
1.	Total params	14,714,688	54,964
2.	Trainable params	14,714,688	53,492
3.	Non-trainable params	0	1,472
4.	Training time	16 mins approx	4 hrs approx
5.	Size of model on disk	211 MB	856 KB
6.	Overall accuracy	64%	87%

Even though MiniXception CNN takes a longer training time it gives accuracy of 87%, it takes much lesser space and the prediction time is much faster as compared to others.

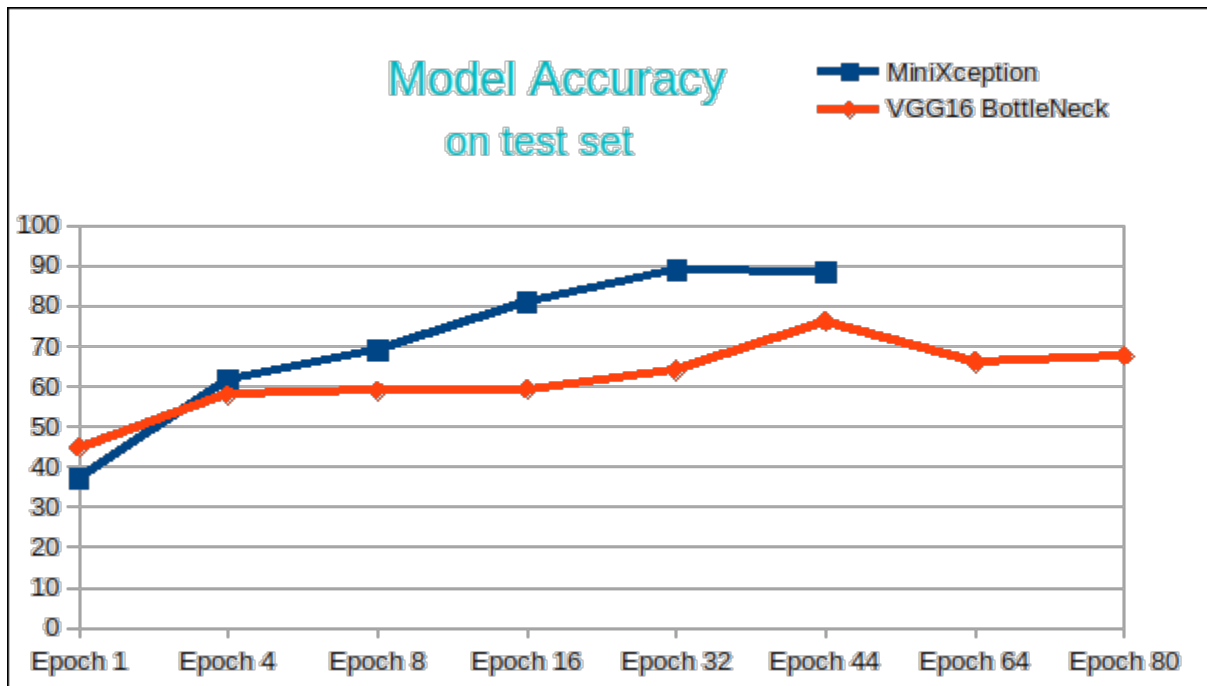
**Comparison of accuracy of different emotion labels:**

	Emotion	VGG16 CNN BottleNeck	MiniXception CNN
1.	Angry	0.62	0.82
2.	Happy	0.78	1.0
3.	Neutral	0.65	0.89
4.	Sad	0.53	0.78



( Fig 15 : ACCURACY COMPARISON)

**Comparison of model accuracy while training :**



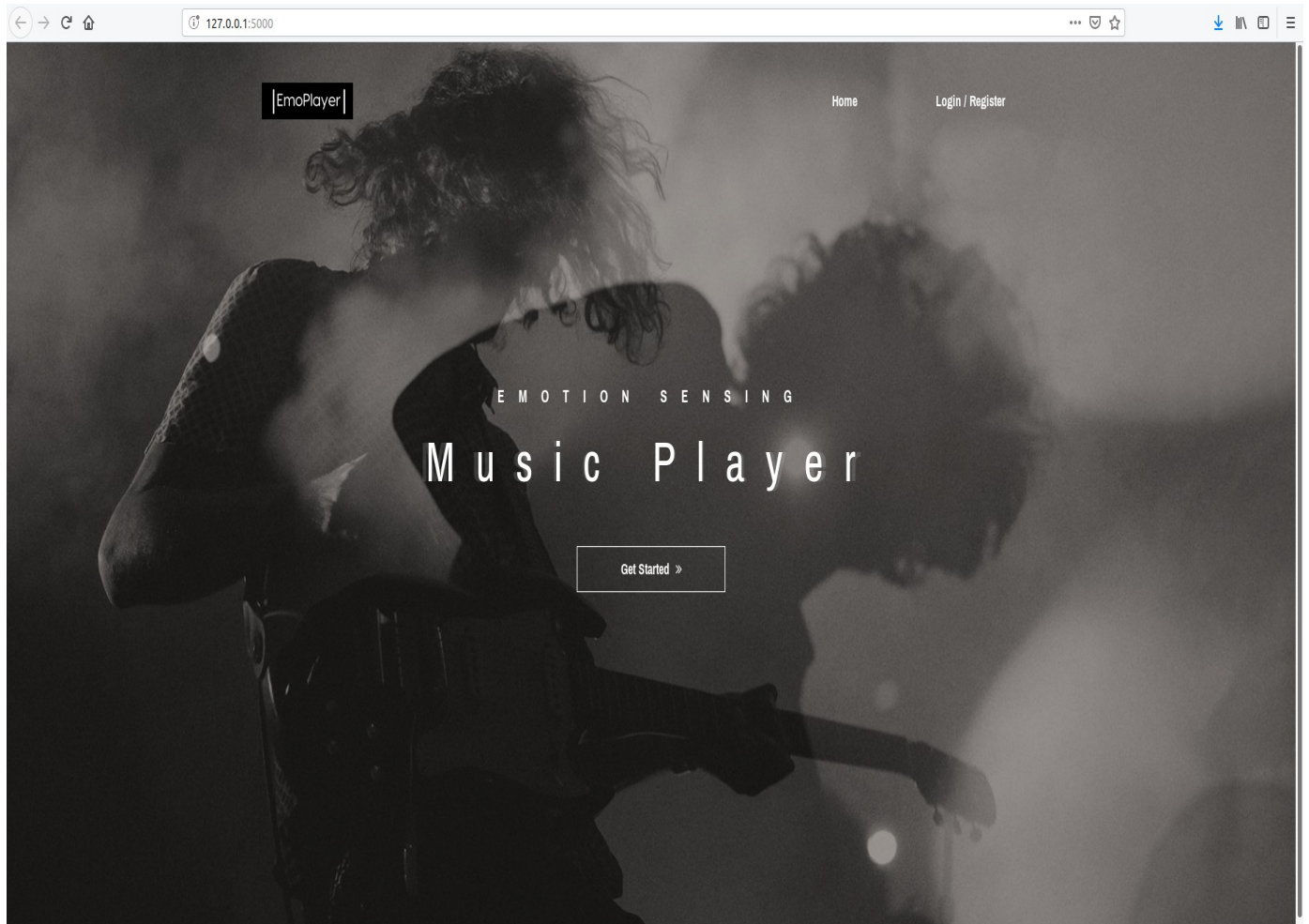
Above diagram shows accuracy of MiniXception and VGG16 BottleNeck when they are trained with different values of epoch.

### Confusion Matrix of MiniXception

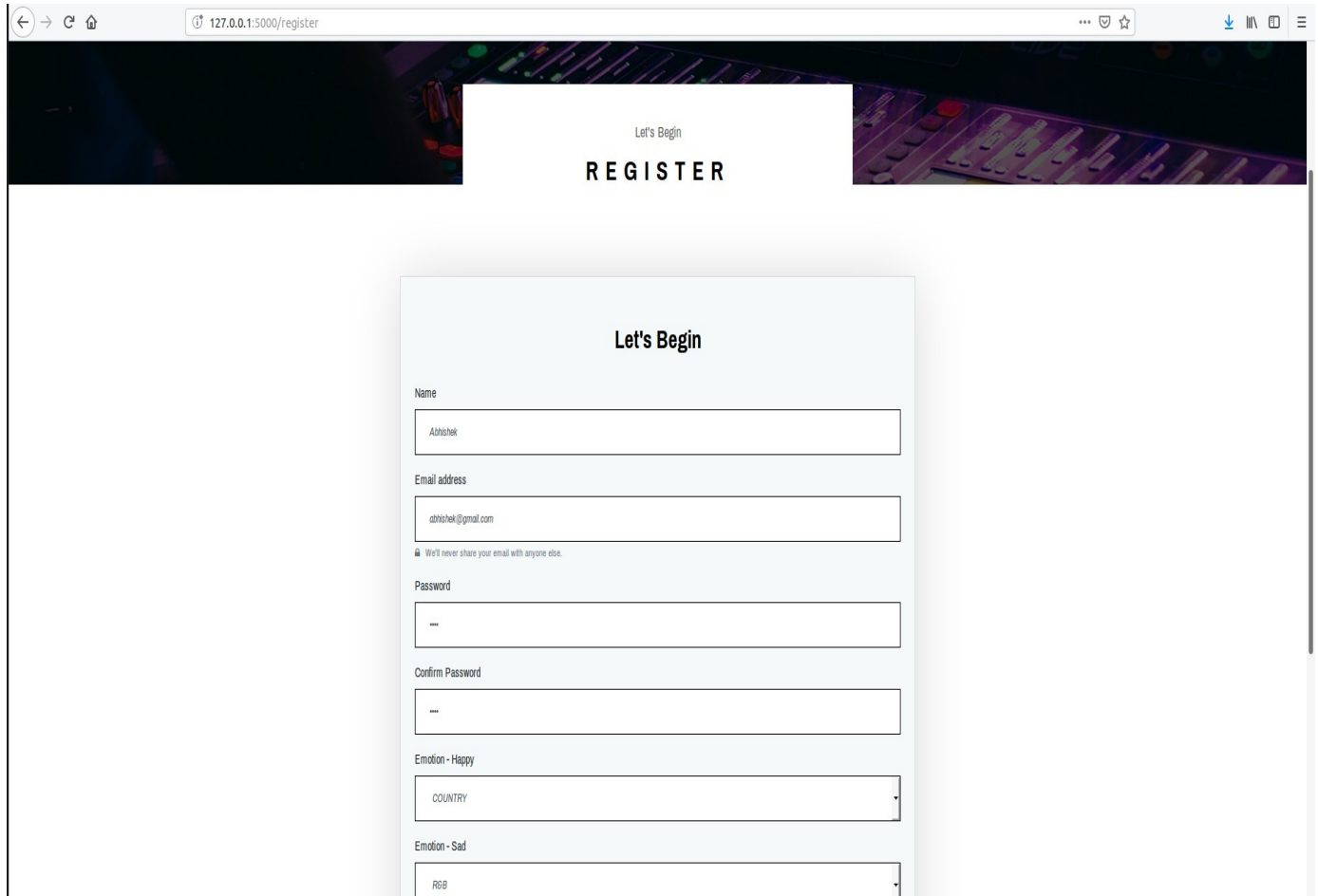
	Anger	Happy	Neutral	Sad	
Anger	880	0	216	126	1222
Happy	0	1144	0	0	1144
Neutral	130	0	980	100	1210
Sad	162	0	279	751	1192
	1172	1144	1475	977	4768

## 6.4 Screenshots

### 6.4.1. Homepage



## 6.4.2. Register



The screenshot shows a web browser window with the address bar displaying "127.0.0.1:5000/register". The page has a dark header with a keyboard background. In the center, there's a white box with the text "Let's Begin" and "REGISTER". Below this, a light blue registration form is displayed. The form includes fields for Name, Email address, Password, Confirm Password, and two dropdown menus for "Emotion - Happy" and "Emotion - Sad".

Let's Begin

REGISTER

Let's Begin

Name

Abhishek

Email address

abhishek@gmail.com

🔒 We'll never share your email with anyone else.

Password

\*\*\*\*

Confirm Password

\*\*\*\*

Emotion - Happy

COUNTRY

Emotion - Sad

RGB

### 6.4.3. Login

See what's new

## LOGIN

### Welcome Back

Email address

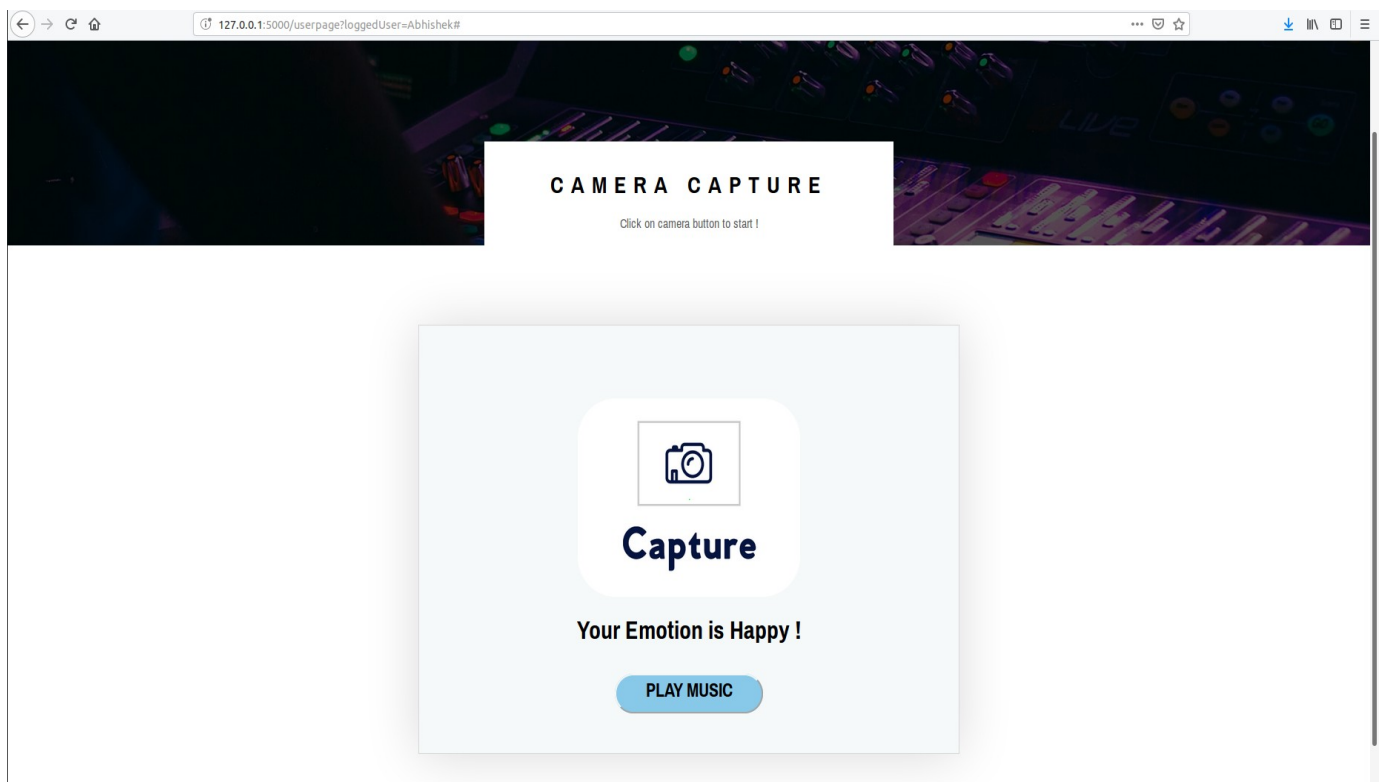
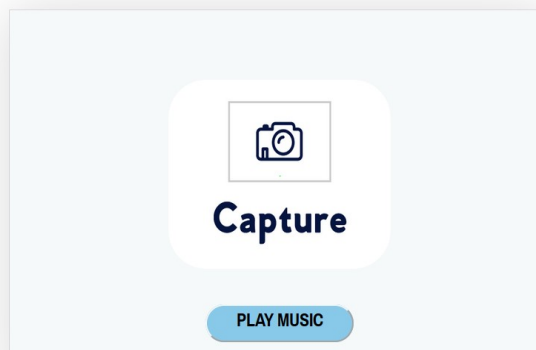
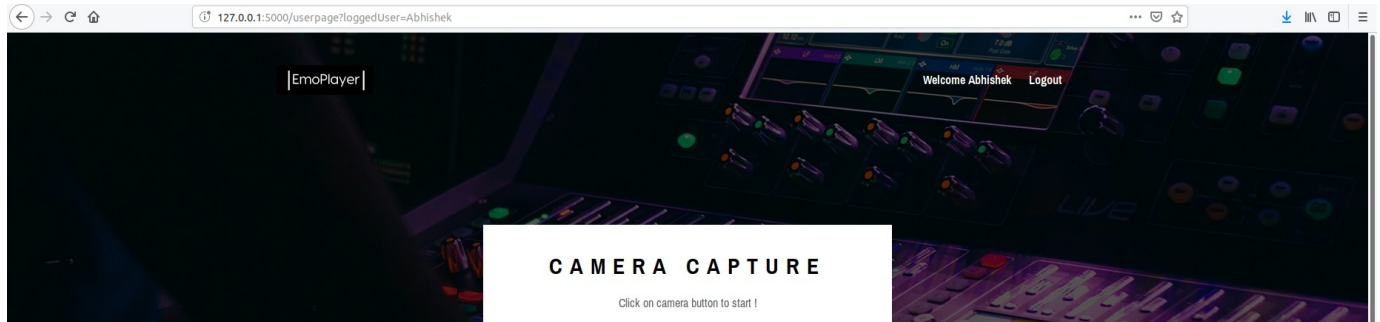
🔒 We'll never share your email with anyone else.

Password

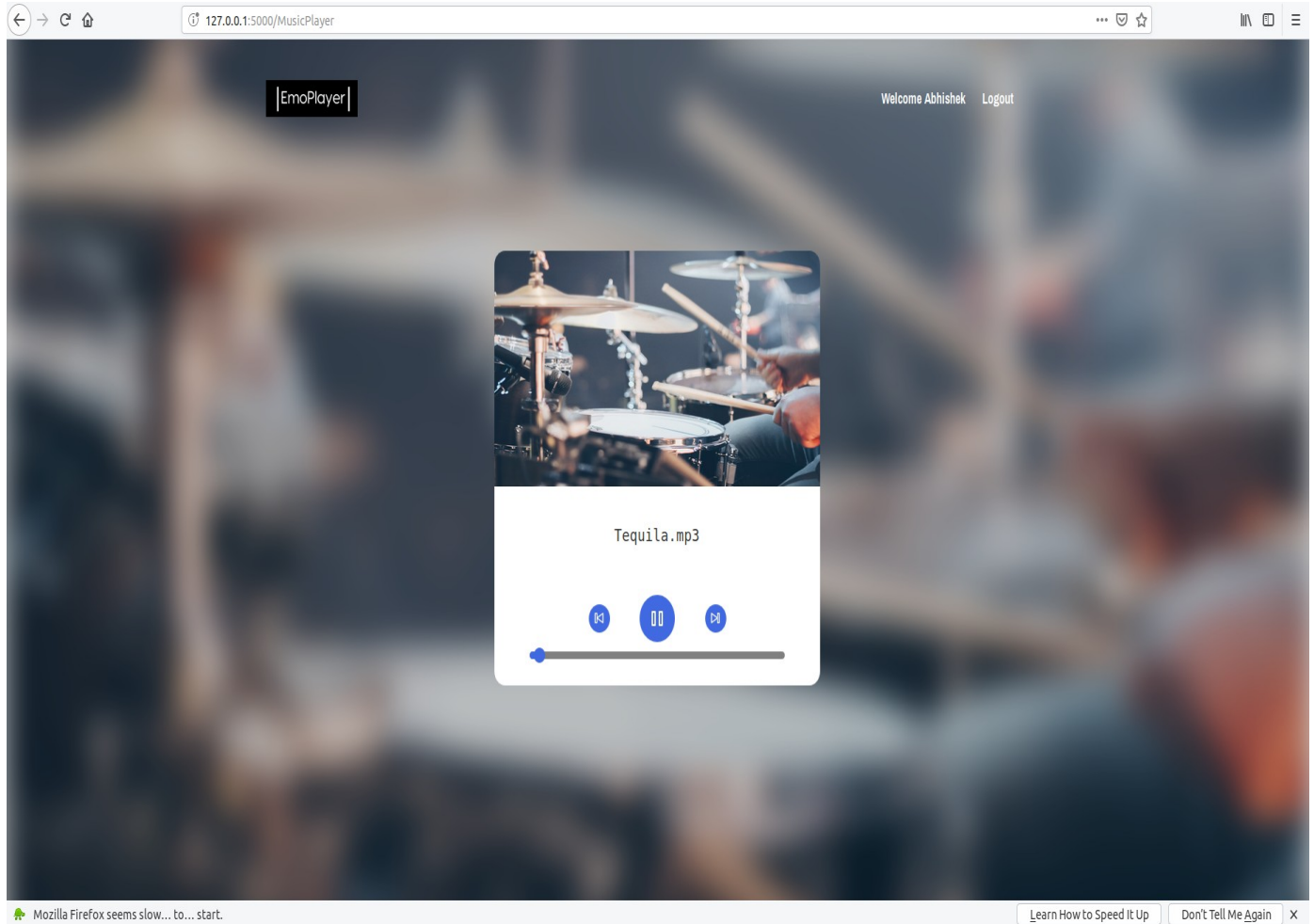
Login

Register

## 6.4.4. Face Capture



## 6.4.5. Music Player



(FIG. 16 Screenshots of Web Application)



## 7. CONCLUSION AND FUTURE WORK

### 7.1 Conclusion:

During the development of the project, existing techniques for emotion recognition were thoroughly researched, highlighting the benefits and problems with each associated. After research MiniXception model was decided. Pros of this model are:

1. Time required for computation is less.
2. MiniXception gives better results as compared to previously ran models.(ref 6.3)
3. The computational time taken is 3 sec which is very less thus helping in achieving a better real time performance and efficiency.

The evaluation of the project indicates that the preliminary objectives of the project have been met. The key objectives were:

1. Create a system that captures user's face and detects emotion.
2. The appropriate music playlists will be played based on user preferences.
3. The system thus aims at providing the web based application system which gives users a cheaper , hardware free and accurate emotion based music system.
4. Reducing the searching time for music thereby reducing the unnecessary computational time and thereby increasing the overall accuracy and efficiency of the system.

### 7.2 Future Scope

- Link the web based application to music streaming services to give user access to large playlists of music.
- Convert the web based application to an android app.

- To design a mechanism that would be helpful in music therapy treatment and provide the music therapist the help needed to treat the patients suffering from disorders like mental stress, anxiety, acute depression and trauma.
- The proposed system also tends to avoid in future the unpredictable results produced in extreme bad light conditions and very poor camera resolution.

# PLAGIARISM REPORT



## PLAGIARISM SCAN REPORT

<b>Date</b>	May 29, 2019	<b>Words</b>	7353
<b>Exclude URL:</b>			
93% Unique		7% Plagiarized	

[Content Checked for Plagiarism:](#)

## REFERENCES

- [1] Martinez, Brais & Valstar, Michel. (2016). Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition. 10.1007/978-3-319-25958-1\_4.
- [2] Anima Majumder, Laxmidhar Behera, Senior Member, IEEE, and Venkatesh K. Subramanian Automatic Facial Expression Recognition System Using Deep Network Based Data Fusion IEEE TRANSACTIONS ON CYBERNETICS
- [3] Lianghua He ,Cairong Zou ,Li Zhao,Die Hu An Enhanced LBP Feature Based on Facial Expression Recognition Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September 1-4, 2005.
- [4] David Charte, Francisco Charte, Salvador Garcia, Maria J. del Jesus and Francisco Herrera, A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. Article in Information Fusion 44:78-96 November 2017
- [5] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," *2017 International Conference on Engineering and Technology (ICET)*, Antalya, 2017, pp. 1-6. doi: 10.1109/ICEngTechnol.2017.8308186
- [6] Cahit Deniz GÜRKAYNAK, Nafiz ARICA, "A Case Study on Transfer Learning in Convolutional Neural Networks", 2018 26th Signal Processing and Communications Applications Conference (SIU), Publisher IEEE
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition", 10 Dec 2015 Cornell University

- [8] Łukasz Kaiser, Aidan N. Gomez, François Chollet, “Depthwise Separable Convolutions for NeuralMachine Translation”, 16 June 2017 Google Brain and University of Toronto
- [9] Ojala, T., M., Pietikäinen and D. Harwood, “A comparative study of texture measures with classification based on feature distributions” Pattern Recognition vol. 29, 1996.
- [10] Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24(7), 971–987 (2002)
- [11] P.Rajeswari,M.G.Sumithra, Survey : Pre Processing Techniques For Facial Expression Recognition
- [12] <http://studentnet.cs.manchester.ac.uk/resources/library/umer.iftikhar.pdf>
- [13] <https://medium.com/@14prakash/transfer-learning-using-keras-d804b2e04ef8>
- [14] <https://medium.com/@sidereal/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5>
- [15] <https://towardsdatascience.com/a-basic-introduction-to-separable-convolutions-b99ec3102728>
- [16] <https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148>
- [17] <http://cs231n.github.io/understanding-cnn/>
- [18] <https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html>
- [19] <https://www.hackerearth.com/practice/machine-learning/transfer-learning/transfer-learning-intro/tutorial/>

