EasyChair Preprint

№ 10377

Mood-Based Emotional Analysis for Music Recommendation

Roshani Raut and Dhruv Goel

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

June 11, 2023

Mood-Based Emotional Analysis For Music Recommendation

Roshani Raut[1] and Dhruv Goel[2]

[1] Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Nigdi,
Pune 411044, India
roshani.raut@pccoepune.org
[2] Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Nigdi,
Pune 411044, India
dhruvgoel01@gmail.com

Abstract. Music plays a crucial role in enhancing mood and motivation, contributing to overall well-being. Recent studies have emphasized the profound impact of music on human brain activity, with individuals exhibiting positive responses to music stimuli. Considering this, people increasingly rely on music to align with their emotional states and personal preferences. This research focuses on a system that employs computer vision techniques to recommend songs based on the user's mood. By analyzing facial expressions, the system accurately identifies the user's emotional state, facilitating an automated music selection process. This approach not only saves significant time but also eliminates the need for manual song browsing. The proposed system eliminates the necessity of human intervention traditionally associated with mood-based music selection, leveraging computer vision technology to automate the process. By employing algorithms such as Haar Cascade and CNN, facial elements are extracted, enabling real-time emotion identification. The use of an internalcamera further optimizes system design efficiency, reducing costs compared to alternative methods. This paper presents an in-depth exploration of the implementation and advantages of an automated mood-based music recommendation system, highlighting the role of facial expression analysis in enhancing user experiences.

Keywords: Convolutional Neural Network, Face Detection, Feature Extraction, Deep Learning, Dataset, Haar-Cascade.

1 INTRODUCTION

Music has long been recognized as a powerful medium that can evoke emotions, uplift moods, and provide motivation. It has the unique ability to deeply resonate with individuals and influence their state of mind. Recent research has shed light on the significant impact of music on human brain activity, revealing the intricate relationship between music and emotions. People increasingly seek music that aligns with their current emotional state and personal preferences, using it as a means to enhance their well-being and overall enjoyment of life. In line with the growing demand for personalized music experiences, this research focuses on the development of an automated mood-based music recommendation system. The system utilizes computer vision techniques to analyse and interpret facial ex-pressions, thereby identifying the user's emotional state in real-time. By under-standing the user's mood, the system can suggest songs that are most likely to resonate with their current emotional state, providing a tailored and enjoyable music listening experience. Traditionally, playing music according to a person's mood required human intervention, such as manually selecting songs or relying on recommendations from others. However, with advancements in computer vi-sion technology, the automation of such systems has become feasible. By lever-aging algorithms like Haar Cascade and Convolutional Neural Networks (CNNs), the proposed system can extract facial elements from live video feeds captured by an internal camera. These facial elements are then analysed to determine the user's emotional state, enabling the system to recommend music that aligns with their mood. One of the key advantages of the proposed system is its ability to significantly reduce the time and effort required for music selection. Instead of manually searching through a vast library of songs or relying on generic mood-based playlists, users can rely on the automated recommendation system to sug-gest music that suits their specific emotional needs at any given moment. This not only enhances user convenience but also ensures a more personalized and im-mersive music listening experience. In this paper, we will delve into the details of the automated mood-based music recommendation system, exploring the under-lying computer vision techniques employed for facial expression analysis. We will discuss the algorithms used, including Haar Cascade and CNNs, and their role in accurately identifying emotions from facial expressions. Additionally, we will highlight the advantages of using an internal camera for facial expression capture, emphasizing the cost-effectiveness and practicality of the system design. Furthermore, we will examine the implications and potential impact of the pro-posed system on user experiences and engagement with music. By providing tai-lored music recommendations based on real-time emotional states, the system has the potential to revolutionize the way people interact with and enjoy music. We will also discuss the future directions and possible enhancements for the system, considering factors such as increased accuracy, expanded music data-bases, and integration with other technologies. Overall, this research aims to ex-plore and demonstrate the capabilities of an automated mood-based music rec-ommendation system that harnesses the power of computer vision and facial ex-pression analysis. By offering a personalized music listening experience, the system aims to enhance mood elevation, motivation, and overall well-being through the power of music.

2 LITERATURE REVIEW

This section provides an overview of the works associated with the proposed system. Each of these studies has adopted unique strategies to address different challenges and advance prior research in the field of emotion recognition. Various approaches for feature extraction and pre-processing are discussed in this section, although our specific research focuses on the CNN model. The intention is to compare the accuracy variations resulting from the integration of these techniques with the CNN model.

Rabie Helaly [06] introduces an intriguing sentiment recognition system, which is implemented on the Raspberry Pi 4 embedded system. Their approach utilizes the Xception CNN model to identify emotions within the system. The embedded system's classifiers categorize captured facial photos into seven distinct facial expressions, which are then used as input for the system. The training process utilizes the FER 2013 dataset. The proposed model achieves a GPU accuracy of 94%. Mohamed Ali Hajjaji, et al. [06] further implemented the system on the Raspberry Pi 4 and achieved an accuracy of 89% while considering GPU limitations.

In addition to the CNN model, Ansamma et al. [07] propose a novel approach to enhance real-time emotion identification. This method incorporates additional feature extraction techniques to improve the accuracy of training. Performance evaluations were conducted using datasets such as FER2013 and JAFFE. The initial module utilizes a webcam for real-time video capture, employing local binary patterns (LBP) to detect faces. The subsequent module focuses on feature selection for emotion recognition and pre- processing. The proposed architecture comprises an input layer, two fully connected classification layers, two pooling layers, and four convolutional layers. The suggested approach demonstrates exceptional performance on the JAFFE and FER2013 datasets, achieving precision rates of 91.2% and 74.4% respectively.

Sabrina Begaj et al. [07] conducted a study on the challenges posed by Emotion Recognition Datasets. In our research, we also explored various CNN configurations and designs. The primary dataset selected by Ali Osman Topal et al. [07] was ICV MEFED. Our deep learning approach involved three main phases: deep feature learning, deep feature classification, and pre-processing. The initial CNN network implemented consisted of two fully connected layers, one dropout layer, four max pooling layers, and four convolutional layers. Upon analyzing the Confusion Matrix, it was observed that the system performed best in detecting pleased expressions while struggling the most in detecting contempt. When evaluated on the FER2013 dataset, the CNN achieved an accuracy of 91.62%. Shuang Liu [05] employed a Convolutional Neural Network (CNN) in conjunction with the Keras deep learning framework. The utilization of Keras proved advantageous due to its user-friendly and modular nature. The key parameters of the convolutional layer include the size of the convolutional kernel, the stride for convolutional translation, and the padding mode. The pooling layer, also known as the subsampling layer, facilitates the down-sampling process. The integration of network layer modules in the design reduces development costs and facilitates easy debugging of the code. The CNN architecture in this paper comprises a total of nine layers, including one input layer and one output layer. Four convolutional layers and pooling layers are employed in the network design.

The author introduces a method that combines CNN with data augmentation, where the image is subjected to pre- processing if a face is detected. The pre-processing step utilizes the Cascade Classifier for face detection. Subsequently, data augmentation is performed using the ImageDataGenerator function available in the Keras API. A standard Facial Emotion Recognition (FER) system comprises three main components:

facial detection, feature extraction, and facial expression categorization. The pre-processing phase, also known as face detection, focuses on identifying facial regions within the images. Facial feature extraction aims to capture the most precise representation of facial images for recognition purposes. In alignment with Phavish Babajee [08], our approach is based on a geometric feature-based method that utilizes an edge detection framework for data extraction.

3 METHODOLOGY

Facial emotion recognition is a challenging task in computer vision and artificial intelligence, and our project aims to develop a robust system using a combination of Convolutional Neural Network (CNN) and Haar Cascade algorithm. In this methodology, we outline the steps involved in training and deploying the model for accurate emotion detection.

The first step is to gather a diverse dataset that consists of facial images with labeled emotions. This dataset serves as the foundation for training the CNN model. Additionally, the Haar Cascade algorithm is employed for face detection and extraction from the images. The algorithm analyzes patterns and features of the detected faces, providing a basis for subsequent emotion recognition.

Once the dataset is collected, we preprocess the images by resizing them to a standard size and applying normalization techniques. This ensures consistency and enhances the quality of input data, which is crucial for the CNN model's training process.

The CNN model's architecture is designed to learn and extract meaningful features from the pre-processed facial images. It typically consists of multiple convolutional layers that detect spatial patterns and features, followed by pooling layers to reduce dimensionality and extract important information. Fully connected layers are then utilized for classification, mapping the extracted features to different emotion classes.

The training phase involves feeding the pre-processed dataset into the CNN model. Through an iterative process, the model learns to identify the underlying patterns and correlations between facial expressions and corresponding emotions. This learning process is guided by optimizing the model's parameters using techniques such as backpropagation and gradient descent.

Once the CNN model is trained, we evaluate its performance using a separate validation dataset. Various metrics such as accuracy, precision, recall, and F1-score are calculated to assess the model's ability to correctly classify emotions. This evaluation step helps us gauge the model's performance and make necessary adjustments if needed.

In the deployment phase, the trained CNN model is integrated with the Haar Cascade algorithm for real-time emotion detection. The Haar Cascade algorithm is responsible for detecting faces in video or image streams, and the CNN model analyses the extracted faces to recognize the associated emotions. This combined approach enables efficient and accurate emotion recognition in real-world scenarios.

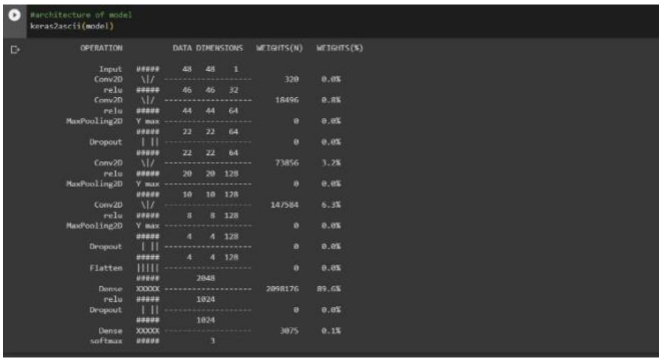
Throughout the project, we utilize project management tools to track progress, manage tasks, and facilitate effective communication within the team. Regular meetings are

conducted to discuss the project's status, address any challenges or obstacles, and ensure smooth coordination among team members.
By leveraging the power of both CNN and Haar Cascade algorithm, our methodology provides a robust framework for facial emotion recognition. The Haar Cascade algorithm contributes to accurate face detection, while the CNN model excels in capturing intricate patterns and features to classify emotions. This integrated approach holds the potential to enhance various applications, such as emotion-aware systems, virtual assistants, human-computer interaction interfaces, and more.

4 PROPOSED SYSTEM

The methodology for facial emotion recognition using CNN for music recommendation involves several key steps. Firstly, a dataset of facial images repre-senting various emotional expressions is collected and reprocessed to enhance their quality and standardize their format. Next, the facial images are labelled with their corresponding emotional expressions to create a supervised learning dataset.

Fig 1: CNN architecture



```
#architecture of model
keras2ascii(model)
```

```
           OPERATION           DATA DIMENSIONS   WEIGHTS(N)   WEIGHTS(%)

               Input   #####     48   48    1
              Conv2D    \|/  -------------------       320     0.0%
                relu   #####     46   46   32
              Conv2D    \|/  -------------------     18496     0.8%
                relu   #####     44   44   64
        MaxPooling2D   Y max -------------------         0     0.0%
                       #####     22   22   64
             Dropout    | || -------------------         0     0.0%
                       #####     22   22   64
              Conv2D    \|/  -------------------     73856     3.2%
                relu   #####     20   20  128
        MaxPooling2D   Y max -------------------         0     0.0%
                       #####     10   10  128
              Conv2D    \|/  -------------------    147584     6.3%
                relu   #####      8    8  128
        MaxPooling2D   Y max -------------------         0     0.0%
                       #####      4    4  128
             Dropout    | || -------------------         0     0.0%
                       #####      4    4  128
             Flatten   ||||| -------------------         0     0.0%
                       #####        2048
               Dense   XXXXX -------------------   2098176    89.6%
                relu   #####        1024
             Dropout    | || -------------------         0     0.0%
                       #####        1024
               Dense   XXXXX -------------------      3075     0.1%
             softmax   #####           3
```

The architecture represents a convolutional neural network (CNN) model. break-down of each layer and its functionality:
1. Conv2D Layer (32 filters, kernel size 3x3, ReLU activation):
• This layer performs convolutional operations on the input images with 32 fil-ters of size 3x3.
• The ReLU activation function is applied to introduce non-linearity into the network.
• The input shape for this layer is (48, 48, 1), indicating images of size 48x48 with a single channel (grayscale).
2. Conv2D Layer (64 filters, kernel size 3x3, ReLU activation):
• This layer performs convolutional operations on the previous layer's output with 64 filters of size 3x3.
• Again, the ReLU activation function is applied.
3. MaxPooling2D Layer (pool size 2x2):
• This layer performs max pooling, which reduces the spatial dimensions of the previous layer's output by taking the maximum value within a 2x2 window.
• It helps in downsampling the feature maps and extracting dominant features while reducing computational complexity.
4. Dropout Layer (dropout rate 0.25):
• Dropout is a regularization technique used to prevent overfitting.
• It randomly sets a fraction (0.25 in this case) of the input units to 0 during training, which helps to prevent the model from relying too heavily on specific features.
5. Conv2D Layer (128 filters, kernel size 3x3, ReLU activation):
• Another convolutional layer with 128 filters of size 3x3 is added.
• The ReLU activation function is applied.
6. MaxPooling2D Layer (pool size 2x2):

- Another max pooling layer with a pool size of 2x2 is added to downsample the feature maps further.

7. Conv2D Layer (128 filters, kernel size 3x3, ReLU activation):

- Another convolutional layer with 128 filters of size 3x3 is added.
- ReLU activation is applied.

8. MaxPooling2D Layer (pool size 2x2):

- Another max pooling layer with a pool size of 2x2 is added.

9. Dropout Layer (dropout rate 0.25):

- Another dropout layer with a dropout rate of 0.25 is added for regularization.

10. Flatten Layer:

- This layer flattens the multi-dimensional output from the previous layer into a 1D vector, which can be fed into a fully connected layer.

11. Dense Layer (1024 units, ReLU activation):

- This fully connected layer consists of 1024 units and applies the ReLU activa-tion function.

12. Dropout Layer (dropout rate 0.5):

- Another dropout layer with a dropout rate of 0.5 is added.

13. Dense Layer (3 units, softmax activation):

- The final dense layer consists of 3 units, corresponding to the output classes of the classification problem.
- The softmax activation function is used to obtain probability distributions over the classes, indicating the model's predicted class probabilities.

In summary, this architecture consists of several convolutional layers for feature extraction, max pooling layers for down sampling, dropout layers for regulariza-tion, a flatten layer to transform the output into a 1D vector, and fully connected layers for classification. The model is designed for a problem with three output classes.

A CNN model architecture is designed, comprising convolutional layers for fea-ture extraction, pooling layers for down sampling, and fully connected layers for classification. The model is trained using the labelled dataset, optimizing its parameters through backpropagation and gradient descent. The trained model is then evaluated on a separate test set to assess its performance in recognizing fa-cial emotions. Once the model demonstrates satisfactory performance, a music recommendation algorithm is developed, which takes the predicted emotional states from the facial emotion recognition model as input. The algorithm matches the emotions with suitable music tracks from a predefined music database or streaming service. A user interface is designed to capture facial images from live video input, process them through the facial emotion recognition model, and pro-vide personalized music recommendations based on the detected emotions. User feedback is gathered through studies or surveys to refine and enhance the Sys-tem. Finally, the System is deployed on a suitable platform or device, with ongo-ing monitoring and consideration for future enhancements.

5 DATASET

The FER2013 dataset is a widely used benchmark dataset in the field of facial expression recognition. It consists of approximately 35,887 grayscale images of size 48x48 pixels, depicting facial expressions across seven categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset was created by collecting images from various sources and is diverse in terms of age, gender, and ethnicity. Each image in the dataset has been labelled with the corresponding emotion category, allowing researchers to train and evaluate machine learning models for automatic emotion recognition. The FER2013 dataset has played a crucial role in advancing the development of facial expression recognition algorithms and has been used in numerous research studies to investigate and improve the accuracy and robustness of emotion classification systems.

6 RESULT

In general, the performance of a facial emotion recognition system using a combination of CNN and Haar Cascade algorithm can be evaluated based on metrics like accuracy, precision, recall, and F1-score. These metrics provide insights into the model's ability to correctly classify emotions and its overall performance.

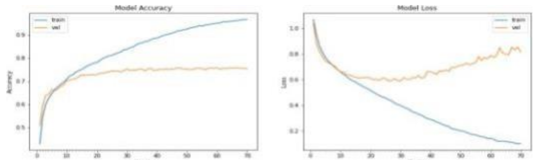


Fig 2 Model Accuracy & Loss

To assess the accuracy of the system, we can compare the predicted emotions with the ground truth labels from a validation or test dataset. Precision measures the proportion of correctly predicted positive emotions out of all predicted positive emotions, while recall measures the proportion of correctly predicted positive emotions out of all actual positive emotions. F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

A confusion matrix is a performance evaluation metric commonly used in machine learning and classification tasks. It provides a detailed summary of the predictions made by a classification model, comparing them to the actual ground truth values. In the context of facial emotion recognition, a confusion matrix can be used to evaluate the performance of an emotion classification model. Each row in the matrix corresponds to a true emotion class, while each column corresponds to a predicted emotion class. The matrix elements represent the number of instances or the proportion of instances that belong to a particular combination of true and predicted emotions.

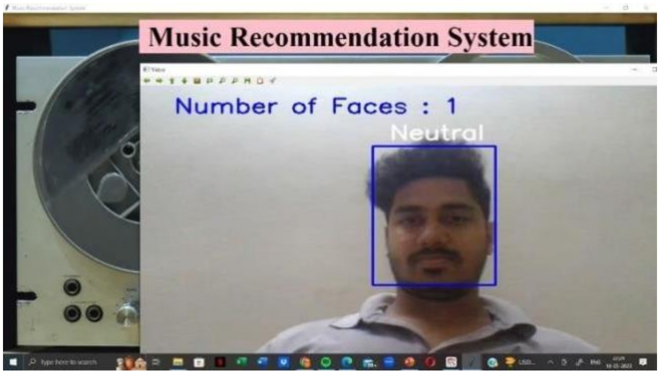


Fig 3: Neutral Emotion Detection

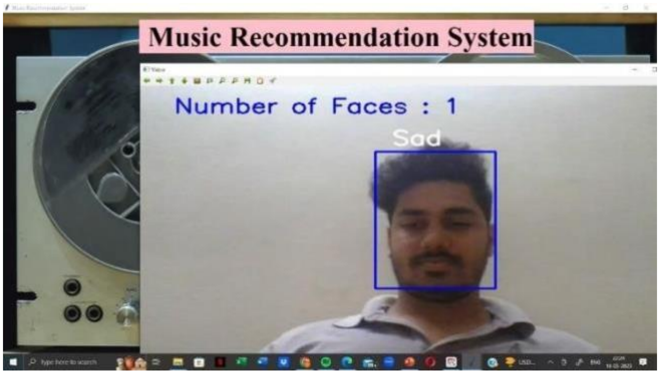


Fig 4: Sad Emotion Detection

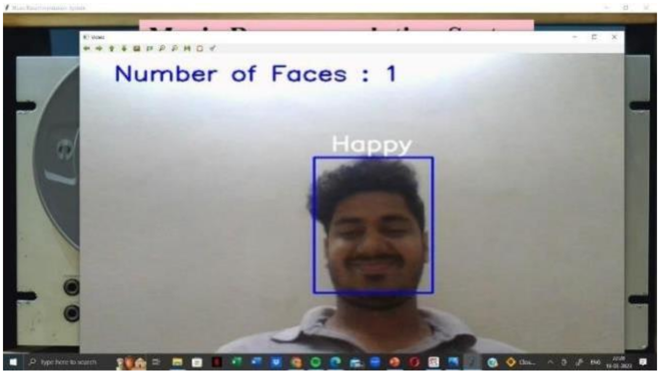
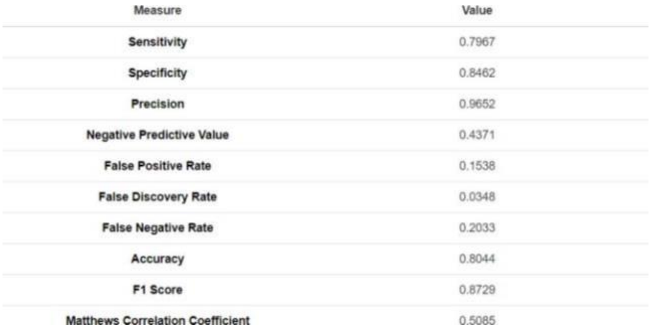


Fig 5: Happy Emotion Detection

Various performance metrics can be derived from the confusion matrix, such as accuracy, precision, recall, and F1- score. These metrics provide a quantitative assessment of the model's ability to correctly classify emotions. Additionally, the confusion matrix can help identify specific emotions that are more challenging to predict accurately and guide further improvements in the classification model. Overall, the confusion matrix serves as a valuable tool for evaluating and analyzing the performance of a facial emotion recognition model, providing insights into the strengths and weaknesses of the model and aiding in the refinement of the classification algorithm.



Measure	Value
Sensitivity	0.7967
Specificity	0.8462
Precision	0.9652
Negative Predictive Value	0.4371
False Positive Rate	0.1538
False Discovery Rate	0.0348
False Negative Rate	0.2033
Accuracy	0.8044
F1 Score	0.8729
Matthews Correlation Coefficient	0.5085

Fig 6: Evaluation Matrix

Conclusion

After conducting a thorough review of existing literature, it became clear that there are many approaches to implementing a Music Recommender System. Drawing upon previous researchers and developers' proposed methods, our team established our System's objectives. WithAI-powered applications' growing popularity and benefits, our project will leverage state-of-the-art technology. OurSystem aims to provide users with Music that matches their mood, considering how Music can impact emotional states. Through the use of facial emotion recognition, our System can detect the user's current Emotion, such as happy, sad, angry, neutral, or surprised. Once the user's Emotion is identified, the System will generate a playlist containingMusic that matches the detected mood. The development process of this application poses challenges due to the intensive CPU and memory usage required to process large datasets. Our goal is to develop the application most cost effectively and standardize it for use on various devices. By using facial emotion recognition in our musicrecommendation system, we hope to simplify creating and managing playlists for users.

References

1. Tawsin Uddin Ahmed, Sazzad Hossain, Mohammad Shahadat Hossain, Raihan Ul Islam, Karl Andersson. Facial Expression Recognition using Convolutional Neural Network with Data Augmentation. 3rd International Conference on Imaging, Vision & Pattern Recognition. 2019 (pp. 336-341).

2. Ramzi Guetari, Aladine Chetouani, Hedi TABIA, Nawres KHALIFA. Real-time emotion recognition in video streams, using B-CNN and F-CNN. 5th International Conference on Advanced Technologies for Signal and Image Processing.2020
3. Ansamma John, Abhishek MC, Ananthu S Ajayan. Real-Time Facial Emotion Recognition System with Improved Preprocessing and Feature Extraction. IEEE Xplore .2020 (pp.1328-1333).
4. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. COMMUNICATIONS OF THE ACM. 2017. (pp. 84- 90).
5. Shuang Liu, Dahua Li, Qiang Gao, Yu Song, Facial Emotion Recognition Based on CNN. Chinese Automation Congress (CAC). 2020 (pp. 398-403).
6. Rabie Helaly, Mohamed Ali Hajjaji, Faouzi Sahli, Abdellatif Mtibaa. Deep Convolution Neural Network Implementation for Emotion Recognition System. IEEE Xplore. 20th international conference on Sciences and Techniques of Automatic Control & computer engineering.2021 (pp. 261-265).
7. Sabrina Begaj , Ali Osman Topal , Maaruf Ali .Emotion Recognition Based on Facial Expressions Using ConvolutionalNeural Network.IEEE Xplore. Central MichiganUniversity. 2021 (pp. 58-63)
8. Phavish Babajee, Geerish Suddul, Sandhya Armoogum, Ravi Foogooa. Identifying Human Emotions from Facial Expressions with Deep Learning. Zooming Innovation in Consumer Technologies Conference (ZINC). IEEE Xplore. 2020 (pp. 36- 39).
9. Akriti Jaisal, A. Krishnama Raju, Suman Deb. Facial Emotion Detection Using Deep Learning. International Conference for Emerging Technology (INCET). 2020 (pp. 1-5).
10. Lyons, Michael; Kamachi, Miyuki; Gyoba, Jiro. The Japanese Female FacialExpression (JAFFE) Database. Zenodo.https://zenodo.org/record/3451524#.Y2paYHZBy3 A
11. Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar Iain Matthews. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion- specified expression. COMPUTER VISION AND PATTERN RECOGNITION IEEE. (pp. 94-101).