Feel Good AI: Voice-Enabled Emotion-based Music Recommendation System

1st Anjali Sharma
Dept. of Computer Science and Engg.
JSS Academy of Technical Education
Bangalore, India
anjali2505sharma@gmail.com

2nd Shubham Vishwakarma
Dept. of Civil and Engg.
Thakur College of Engineering and Technology
Mumbai, India
vishwakarmashubham.2503@gmail.com

3rd Liya T Mathew
Dept. of Electronics and Communication Engg.
Cochin University of Science and Technology
Kochi, India
liyatmathew23@gmail.com

***Abstract*—Feel Good AI is a novel effort committed to improving psychological well-being by means of methodically providing customized music suggestions that are in accordance with the user's emotional state. By combining sophisticated recommendation algorithms with advanced technologies, most notably Convolutional Neural Networks (CNN) for facial emotion recognition, this project effortlessly develops innovative tech and empathetic emotional support. The users interact with an intuitive chatbot that generates unique music suggestions through the use of personalized conversations in which they are asked specific emotion-related inquiries. Beyond immediate relevance, incorporating AI-powered music curation to enhance positive emotional experiences for individuals, particularly those afflicted with dementia, is of greater significance. We live in a society where stress and mental health problems are prevalent. By combining problem-solving and emotional well-being, Feel Good AI offers evidence that technology can enhance the quality of life.**

***Keywords*—CNN, Facial Emotion Recognition, Voice Assistant, Customized Music Suggestions**

I. Introduction

In the growing climate of technology-driven solutions for mental well-being, our efforts offer an innovative approach to matching artificial intelligence and music therapy. The primary objective of this study is to develop a music recommendation system that utilizes sophisticated algorithms and a personalized chatbot interface, with a specific emphasis on emotions. As the importance of tension and emotional health is increasingly recognized as vital components of holistic well-being in contemporary society, our efforts strive to meet a fundamental human need—the desire for a private space for personal listening. Our system goes beyond traditional music devices by incorporating a chatbot and an emotion capture camera for micro-interactions. The playlist is then seamlessly curated by the underlying algorithms, tailored to the user's emotional profile. With so many technical solutions available,

Our research sets it apart by acknowledging the emotional resonance of music. This paper delves into emotion-fueled music recommendations, highlighting the role of chatbots in enhancing user interaction. By offering this new approach, we contribute to the discourse around technology, emotions, and mental health. Driven by the vision of enhancing well-being through technology, our research cuts across artificial intel- ligence, emotion recognition, and music therapy. Revealing the complexipersonalizedion-based music recommendation system, this paper not only presents a technical solution but also seeks to provide a discussion of the evolving state of AI in mental health support encouraged as well.

The inception of "Feel Good AI," a voice-enabled emotion-based music recommendation system, marks a significant stride towards amalgamating technology with mental well-being. This system, by employing sophisticated recommendation algorithms alongside CNN for facial emotion recognition, offers customized music suggestions aligning with the user's current emotional state. This endeavor not only addresses the immediate need for personal emotional support but also broadens the horizons of AI in mental health applications.

II. Literature Review

Integrating human emotion recognition with music classification has been a focus of research, recognizing the importance of facial expressions in emotional messages and exploring the mechanisms of imaging, facial recognition, and learning processes, such as Artificial Neural Networks (ANNs) and Hidden Markov Models (HMMs) are used. Large datasets such as the Cohn Kanade dataset help to train these models. In parallel, research in music emotion recognition has evolved, encompassing dataset creation and feature extraction to capture emotional nuances in songs. The Arousal-Valence model proposed by Thayer has been instrumental in determining music emotion, with arousal representing energy and valence indicating stress. Music is categorized based on these factors, forming quadrants representing various emotions. Noteworthy contributions in literature include H. Zhang, K. Zhang, and N. Bryan-Kinns' exploration of music's emotional role in daily activities for context-aware recommendations. Their Emo-Music service, validated through user studies, excels in recommending emotionally tailored music based on daily activities, emphasizing cross-cultural perspectives.[1] Yu-Hao Chin et al.'s work introduces an emotion profile- based music recommendation system utilizing support vector machines (SVM). Emotion profiles, derived from short-term and long-term features, combine with recognized emotions and historical queries for personalised recommendations, showcas- ing an integrated approach to emotion content detection and SVM-

based emotion recognition.[2] Yoon et al.'s research focuses on a Music Recommendation System incorporating low-level features tied to human emo- tions, utilising audience ratings for personalised recommenda- tions. Their findings affirm the effectiveness of low-level fea- tures in predicting emotional responses and enhancing music recommendations, contributing valuable insights to emotion- based music analysis.[3] Shlok Gilda and colleagues address challenges in emotion-based music classification and recommendation, introducing the EMP music player. Combining facial emotion recognition and audio features with deep learning algorithms yields high accuracy in sentiment-aware playlist generation, providing insights into cost-effective and efficient emotion-aware music systems.[4] Chankuptarat et al.'s Emotion-Based Music Player excels in recommending personalised songs based on users' emo- tional states, integrating heart rate and facial image analysis. Notably accurate in identifying happy emotions, their dual classification method exemplifies the promising integration of physiological and visual cues in emotion-aware technology.[5] S Metilda Florence and M Uma's(2020) study proposes an innovative approach to emotion-based music recommen- dation by leveraging user facial expressions. Experimental results showcase the system's ability to predict user emotions and recommend music with high accuracy, prompting future improvements, including reduced classifier training time and exploration of EEG signals for enhanced emotion detection.[6]

III. Methodology

A. Data Collection

The Jonathan Oheix image dataset, sourced from Kaggle and contributed by Jonathan Oheix, comprises two separate folders specifically allocated for training and testing. The training set consists of 28,273 images, with the following distribution across classes: Angry (3,993), Happy (7,164), Neutral (4,982), Fear (4,103), Sad (4,938), and Surprise (3,205) as shown in fig 1. The testing set comprises a total of 7,067 images, which are classified into the following categories: Angry (960), Happy (1,825), Neutral (1,216), Fear (1,018), Sad (1,139), and Surprise (797).[7] The dataset comprises approximately 35 thousand files, occupying ap- proximately 58 MB of storage space. Tanul singh's Twitter dataset, which was contributed by him and was obtained from Kaggle, is a CSV file containing Twitter view data with 162980 rows of information for sentiment analysis. Each folder based on emotion contains 20 tracks for recommendations music. Some of the images has been removed to standardized the Collection of data for each emotions folder and finally founded dataset size to be around 3000 for each label as shown in fig 2. This was done to make my model avoid over-fitting state and bias decision made by machine learning model. But removal of dataset make model train data redundancy so to cope-up this it is good practice to always have more data in dataset.

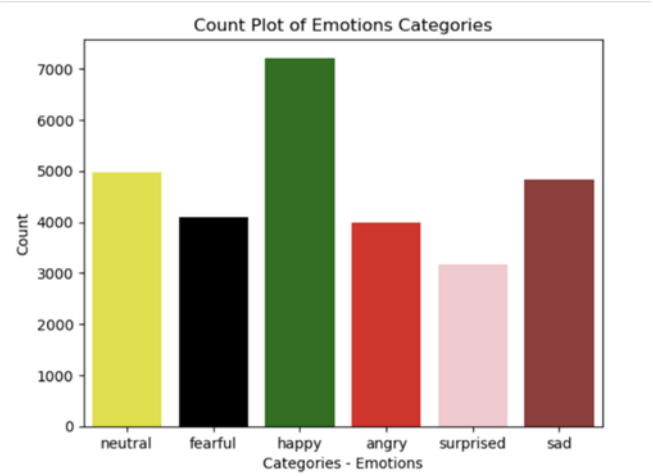

Fig. 1. Original number of image in each label folder

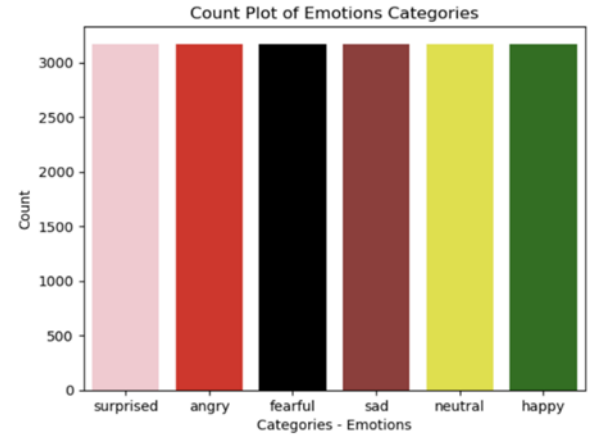

Fig. 2. Final number of image in each label folder

B. Feature Extraction

Feature extraction is an essential step in reducing the dimensionality of data. It involves dividing and condensing the raw data into more manageable groups. This facilitates the handling of extensive data sets containing numerous variables. Feature extraction is the process by which Convolutional Neural Networks (CNNs) identify significant patterns within an image to classify it. The approach employed is object-based, wherein an object is defined as a cluster of pixels sharing common attributes. Conventional classification methods are based on pixels, where the spectral information in each pixel is utilized to classify imagery. The image is converted to grayscale once it has been loaded. As a result, the final image will have only one color channel, which will represent grayscale intensity values. After loading, the image is transformed into a NumPy array, a common practice in image processing. This array is prepared for use in machine learning models, especially for picture classification tasks. Before vectorizing text for sentiment analysis, data was cleaned to ensure proper

English representation.

C. Algorithm

Convolutional Neural Networks (CNNs) are specialized types of neural networks that are adept at detecting and classifying distinct features present in images.[8] They are widely used for the examination of visual images. These applications include tasks such as identifying and categorizing images and videos, analyzing medical images, processing visual information, and understanding human language. CNN exhibits a notable level of precision, rendering it valuable for image recognition purposes. Image recognition is utilized across diverse industries, including medical image analysis, phone technology, security systems, and recommendation systems. In the original mathematical expression, Convolution refers to the adding of two functions to constitute a third function which determines how one of the original two deforms the other. This process involves multiplying the two functions together in a specific way. The method primarily consists of multiplying two matrices, each representing an image, to produce an output. This output is then used for extracting features from the image.

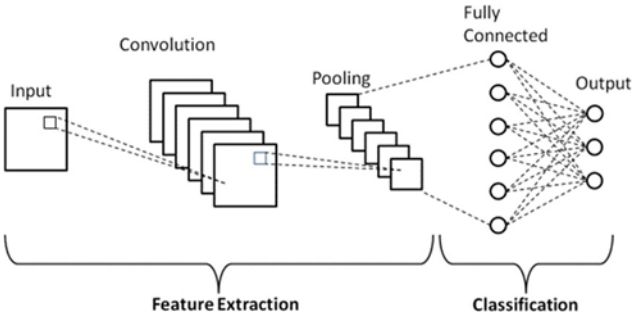

Fig. 3. CNN Architecture

The CNN architecture described in fig 3 is a sequential model constructed using the Keras framework. Below is an analysis of its composition and the number of tiers:

Input Layer: The input shape specified is (48, 48, 1), indicating that this model expects input images of size 48x48 with a single color channel (likely grayscale). Convolutional Layers: The initial convolutional layer of the network comprises four layer CNN with 128, 512 filters, with combination of 3x3 and 5x5 kernel size, and employs a Rectified Linear Unit (ReLU) activation function. The second convolutional layer contains 128 filters, also 5x5 in size, activated by ReLU. The third convolutional layer is equipped with 512 filters, each 3x3, and uses the ReLU activation function, similar to the fourth convolutional layer, which also includes 512 filters of the same size and utilizes ReLU.[9]

The model architecture includes convolutional layers with filter size f, channel size c, stride s, and padding p. Max Pooling layers with a 2x2 pool size follow each convolutional layer, totaling four Max Pooling layers. Additionally, there are seven Dropout layers with a dropout rate of 0.25 after each fully connected layer, convolutional layer, and Max Pooling layer.

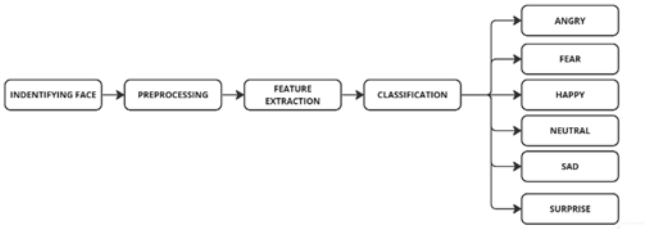

Fig. 4. Face Recognition Architecture.

A Flatten Layer is used to streamline the output from the convolutional layers before forwarding it to the fully connected layers. The network includes two dense layers: the first contains 512 neurons with ReLU activation, and the second comprises 256 neurons, also with ReLU activation. The final output layer is a dense layer with 6 neurons, using the softmax activation function, indicating that the model is designed for a classification task involving six distinct categories as shown in Fig. 4.

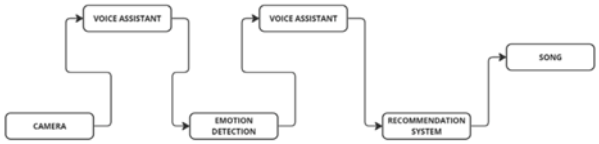

Fig. 5. System Design

The CNN algorithm, built with Keras and TensorFlow in Python, can detect five facial emotions: happiness, sadness, anger, surprise, and neutrality in real time. OpenCV handles face detection in live webcam feeds. Detected faces are processed and classified by the trained neural network for emotion detection. The use of deep learning in recognizing facial expressions significantly diminishes the dependence on face-physics-based models and preprocessing methods, allowing for direct learning from input images and removing the need for intermediary steps in the process. The voice assistant uses pyttsx3 for speech synthesis and speech recognition for understanding spoken commands. Together, these tools enable the assistant to both generate speech and accurately interpret user instructions, providing a seamless and interactive experience. With pyttsx3, the assistant can speak responses, while speech recognition allows it to understand and execute spoken commands effectively, enhancing usability and engagement.

Subsequently, the voice assistant employs a CNN model to predict the user's emotions before engaging in conversation.

Following the interaction, the system analyzes the discourse to determine its tone—positive, negative, or neutral—as an additional measure to ensure accurate emotion detection and assess the user's mood. Based on the detected emotion, the system plays audio or music that corresponds to the expressed

emotion, offering users a selection of songs linked to various emotions as shown in Fig. 5. Users are encouraged to listen to the tracks in the displayed sequence for consistency.

IV. Results

The CNN model achieved 62% accuracy on training data and 53% on unseen test data as mentioned in table 1. It accurately predicts happy, neutral, surprised, and angry emotions 90% of the time. Still, it struggles with distinguishing between sad and fearful, often predicting neutral or angry instead, leading to lower overall accuracy. The sentiment model, with a test accuracy of 89% and train accuracy of 93%, is effective in identifying a person's mood from voice input, aiding in addressing anticipatory feelings of fear and sadness. This ensures that the song suggested to the user is mostly accurate. The project was deployed using Flask.

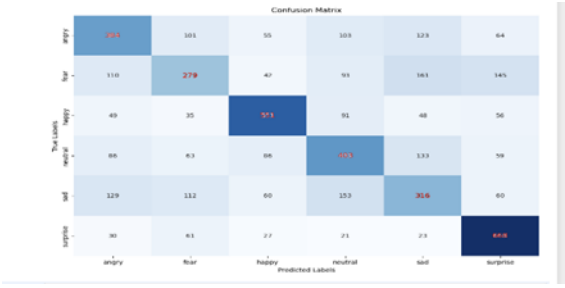

Fig. 6. Validation-Training loss and accuracy

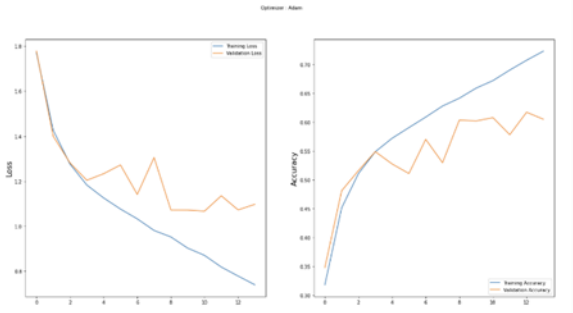

Fig. 7. Validation-Training loss and accuracy

TABLE I
SHOWS THE ACCURACY OF OUR MODELS

MODEL	TRAIN ACCURACY	TEST ACCURACY
Emotion	62	53
Sentimental	93	89

The Sentimental has helped our Emotion model to negotiate incorrect prediction of the labels like fearful and angry by giving the sense of negative sentiments so that at the end opposite songs were recommended by the system.

V. Conclusion

Our project represents a pivotal advancement in the fusion of technology and mental well-being within the realm of music

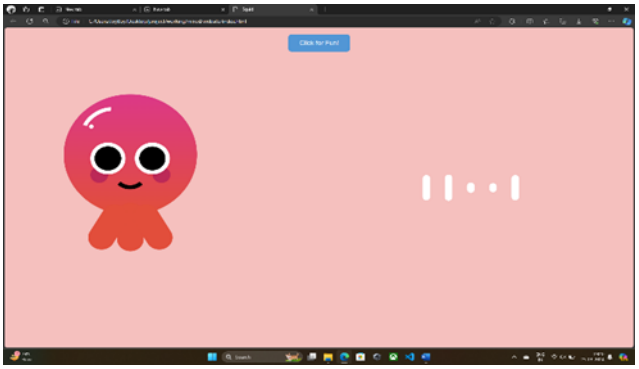

Fig. 8. WebApp display

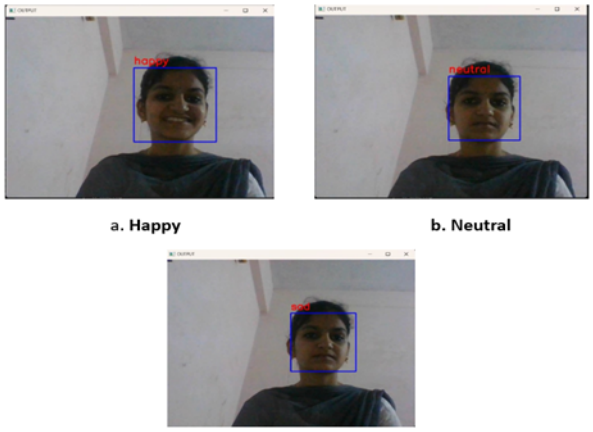

Fig. 9. Shows live emotion detection process in action

streaming. Addressing a critical void in existing platforms, our system delivers real-time, emotion-specific music recommendations to uplift users' moods. By seamlessly integrating emotion-based suggestions with accessible music therapy, our platform emerges as a distinctive and widely accessible solution.

Through precise mood detection and advanced algorithms continually refining datasets, users, benefit from personalized and beneficial music recommendations, showcasing the proven potential to impact emotional states positively.

Positioned as a transformative force, our commitment extends towards incorporating additional parameters like auditory features and physiological markers, surpassing reliance on lexical expressions. Future enhancements, including lyrics analysis, promise a comprehensive understanding of users' emotional states, reinforcing our dedication to technological innovation for mental well-being. As we evolve, we anticipate our project making a significant and positive imprint on the intersection of technology and emotional wellness.

For precise identification of fear and disgust emotions, it is essential to include factors like heart rate and body temperature, rather than relying only on facial expressions. Furthermore, selecting appropriate music to play when these

moods are detected presents a challenge. Therefore, these aspects could be explored as potential future developments for our project.

VI. Future Scope

The "Feel Good AI" project will redefine the present context of emotion-based music recommendations with solutions that will be more empathetic and context-aware and contextual than ever before, through cutting-edge technologies and methodologies. The gist of such an innovation would be to imbed the multimodal techniques for recognizing emotions, wherein the analysis of the facial expression by the user is added to the finer details of voice tonality and physiological signals such as heartbeat and body temperature to get a whole picture of the emotions expressed. This holistic approach is poised to enhance the accuracy of music suggestions significantly. With that, he explained that the project seeks to enhance the user experience with contextual awareness. That means suggestions, besides being timely, will be based on factors like location, time of the day, and current activity to make them perfectly fitting for his immediate environment and intended activities.

The service uses machine learning algorithms designed to adapt to every interaction between the service and a user to give a truly personalized service. Thus, the system could learn with dynamism from prior preferences and feedback in such a way that over some time, it could better predict needs. At the same time, the development of the music database itself that will include expanded ground in genres and sub-genres, plus advanced categorization by mood, tempo, and lyrical content sounds like an ability to offer diversified musical selections in the end. The use of music analysis techniques to understand the emotional and thematic content of songs further elaborates the matching of music to a user's emotional state with more usage.

Another fundamental aspect of the project is to work with music therapists and psychologists so that therapeutic findings are integrated into the recommendation algorithms. Such collaboration will make the system better not only in selecting music that appeals to the mood but also useful to increase the therapeutic potentials that music offers toward mental well-being. However, the system probes further into understanding the emotions, thus ethical considerations and user privacy become the most core issues. The project is, however, committed to ensuring that there is strong protection of data and that such processes of user consent remain transparent so that sensitive information is maintained and safeguarded for trust.

Finally, having in mind the great tapestry of global music preferences, and at the same time being aware that cultural differences might influence emotional expression, the project purports to investigate cross-cultural music recommendations. The "Feel Good AI" project seeks to be a system of recommendations that would culturally fit all users worldwide, regardless of their origin. This paper seeks to identify and respect the cultural peculiarities in the perception of music. All this together denotes a great leap in the use of technology, music, and psychology to bring forth a service that not only understands the emotional world of the user but equally enriches it with the healing and transformational power that music brings.

Acknowledgment

The authors wish to express their sincere gratitude to Indamma Gunna for granting permission to use her image in this study. Her willingness to contribute to the advancement of research is greatly appreciated. This gesture has significantly enhanced the quality and authenticity of our work.

References

[1] H. Zhang, K. Zhang and N. Bryan-Kinns,"Exploiting the emotional preference of music for music recommendation in daily activities," 2020 13th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2020, pp. 350-353, doi: 10.1109/ISCID51228.2020.00085.

[2] Y. H. Chin, S. H. Lin, C. H. Lin, E. Siahaan, A. Frisky and J. C. Wang, "Emotion Profile-Based Music Recommendation," 2014 7th International Conference on Ubi-Media Computing and Workshops, Ulaanbaatar, Mongolia, 2014, pp. 111-114, doi: 10.1109/U-MEDIA.2014.32.

[3] K. Yoon, J. Lee and M. -U. Kim, "Music recommendation system using emotion triggering low-level features," in IEEE Transactions on Consumer Electronics, vol. 58, no. 2, pp. 612-618, May 2012, doi: 10.1109/TCE.2012.6227467.

[4] S. Gilda, H. Zafar, C. Soni and K. Waghurdekar, "Smart music player integrating facial emotion recognition and music mood recommendation," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 2017,pp.154-158,doi: 10.1109/WiSPNET.2017.8299738.

[5] K. Chankuptarat, R. Sriwatanaworachai and S. Chotipant, "Emotion-Based Music Player," 2019 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST), Luang Prabang, Laos, 2019, pp. 1-4, doi: 10.1109/ICEAST.2019.8802550.

[6] S Metilda Florence and M Uma 2020 IOP Conf. Ser.: Mater. Sci. Eng. 912 062007

[7] https://www.kaggle.com/jonathanoheix/face- expression- recognition dataset

[8] Mahadik, Ankita Milgir, Shambhavi Patel, Janvi Kavathekar, Vaishali Jagan, Vijaya Bharathi. (2021). Mood based music recommendation system.

[9] Kaur, Prabhjot Kapoor, Aditya Solanki, Yash Singh, Piyush Sehgal, Devansh. (2022). Deep Learning Based Emotion Detection in an Online Class. 10.1109/DELCON54057.2022.9752940.

[10] Rahul ravi, S.V Yadhukrishna, Rajalakshmi, Prithviraj, "A Face Expression Recognition Using CNN and LBP",2020 IEEE.

[11] F. Abdat, C. Maaoui and A. Pruski, "Human computer interaction using emotion recognition from facial expression", 2011 UKSim 5th European Symposium on Computer.

[12] Gupte A, Naganarayanan A and Krishnan M Emotion Based Music Player-XBeats International Journal of Advanced Engineering Research and Science 3 236854

[13] Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z and Matthews I 2010 The extended cohn-kanade dataset (ck+) A complete dataset for action unit and emotion-specified expression In 2010 ieee computer society conference on computer vision and pattern recognition-workshops 94-101 IEEE

[14] S.Deebika, K.A.Indira, Jesline, "A Machine Learning Based Music Player by Detecting Emotions",2019 IEEE.

[15] S.L.Happy and A. Routray,"Automatic facial expression recognition using features of salient facial patches," in IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 1-12, 1 Jan.-March 2015.