Music Recommendation system based on facial emotion expression

Sarthak Singh

Student of Department of Computer Engineering

Mukesh Patel School of Technology Management & Engineering

Mumbai, INDIA

sarthaksingh1211@gmail.com

Abstract—

In the world of picking songs for people, most old ways just look at what you've liked before and what you've heard. They often miss how you feel – a key part. This study brings a new way to suggest songs, using new tech that can tell how you feel by looking at your face in real-time. It uses ConvLSTM and CNN tech. The ConvLSTM can spot your feelings from your face 89% of the time, and the CNN helps pick songs that fit with those feelings with 97% right picks. By watching your face, the system knows how you're feeling and finds songs that match. Linking ConvLSTM for seeing your mood and CNN for choosing songs makes it hit home with your emotions, making you bond more with your music. This fresh idea could make people more into their music, help their minds feel better, and just make listening to music better overall. Looking ahead, we could make the feeling data better, add more stuff to make even better song picks, and check how happy users are to make the system even better.

Index Terms—Movie Recommendation, Sentiment Analysis

I. INTRODUCTION

In today's evolving world of music consumption the land- scape of personalized entertainment and accessibility has un- dergone significant changes. The vast array of music available, through streaming services, coupled with the preferences of users has created a growing demand for recommendation al- gorithms that break boundaries. While current algorithms have made advancements by combining user preferences, listening history and demographic information there is an aspect of the music experience that often gets overlooked; emotions. [3], [14]

Acknowledging the significance of connection in music, the forefront of recommendation system development is embark- ing on an exploration of new avenues. In addition to metrics emerging algorithms are delving into the realm of emotions by incorporating technologies like real time facial expression monitoring during listening sessions. This groundbreaking approach aims to understand the nuanced responses evoked by music through the cues found in facial expressions.

These thinking recommendation algorithms strive to en- hance the accuracy and depth of music suggestions by har- nessing face recognition technology to gain insights into the listeners state. This distinctive method goes beyond analyzing preferences; it seeks to tap into the visceral and, at times unspoken emotional reactions that music can evoke. The main goal is to provide customers with a captivating and emotionally engaging journey where the suggestions not align with their preferences but also resonate on a profound emotional level. [10]

In short, the changing landscape of music recommendation systems is not only about refining personalized playlists based on known listening behaviors, but about the building of a more human-like platform that can learn and respond to the multifaceted feelings that weave their way into the very texture of musical experiences. [6] This breakthrough may lead to a new ergonomics where people not only find the music to match their tastes, but also get to establish a more profound and emotionally engaged relationship with the acoustic ingenuities that denote the auditory creativity of the music they own.

II. METHODOLOGY

A. Data Collection

1) **Facial Emotion Recognition:** The CK+48 dataset can be regarded as a very vital database in the study of facial expression analysis and emotion identification, which expands and revises the previous CK dataset. This piece had been created with care and also attention, and as such, it forms a complex, polychromatic thread of human emotional states, delving into 48 distinct emotion classes. The sessions comprise various emotions like the rapid jubilant and also unsettling ”Flight of the Bumblebee”, the absolutely impeccable slow metaphors for ”Petite Mort” and the emotionless static music for the ”Stati” (”Stasi”, ”The Bored”), and the piteous and sorrowful ”Les Larmes de V

640x490 resolution grayscale faces in the collection are collected under the illumination consistency; therefore, this makes for a clear evaluation environment. This robust vetting process maintains the consistency which, then, makes CK+48 a resource for the scholars and practitioners alike, useful and relevant in their studies. Also, one of the methods used in face recognition which involves the inclusion of photos from several participants makes it more complex because it depicts the unique facial expressions between the individuals.

Besides conducting as a standard, the CK+48 at the same time serves as a catalyst for the ingenuity in facial expression detection systems. Researchers design their models in the light of its rich and also complete nature to check its accuracy, robustness, and also generalization. This dataset is the very vital part of the research landscape, allowing for the devel- opments in the automatic emotion detection systems and also enabling the betterment of human-machine interactions as it advances our comprehension of how the machines understand and respond to the multiple expressions of human emotions. As a monument to it, CK+48 is still a moving force behind the development of advanced algorithms applied to facial expressions, which are to be the part of the growing context called affective computing.

Paper	Summary	Algorithm Used	Results	Pros	Cons	Accuracy
[1]	Combines emotion recognition and music recom- mendation with explain- able AI techniques.	ResNet50	82% accuracy in emotion classifi- cation	High accuracy; explainable AI improves transparency	Might require substantial computational resources	82%
[7]	Detects emotions via fa- cial expressions to recom- mend music.	CNN	High effectiveness in matching music to emotions	Direct interaction through Pygame	Only focuses on a subset of emo- tions	95.14%
[4]	Discusses a music recom- mendation system using facial expression recogni- tion to enhance user expe- rience.	CNN, KNN	High effectiveness in matching music to detected emotions	Uses various ML techniques to improve accuracy and user interaction	Might require high computational resources	62.1%
[5]	Focuses on real-time fa- cial emotion classification using deep learning for se- curity and surveillance.	CNN	Achieved high accuracy in emotion classification	Real-time processing capabilities	Requires sophis- ticated hardware for real-time pro- cessing	88%
[9]	Introduces a two-part CNN model for facial emotion recognition improving accuracy by focusing on background removal and feature extraction.	Two-part CNN	High accuracy in various test con- ditions.	Robust against background noise; uses deep learning effectively.	High computational power needed	96%

2) **Music Recommendation System***:* The Multi-modal Mixed Signals (MIREX) Emotion Dataset is a crucial resource in the exciting world of emotional computing particularly in the division of multi-modal emotion recognition. This collection is associated with the MIREX Music Information Recognition Evaluation

Exchange (MIREX), thus it gives a human-level multi-modal spectrum of emotional expressions. It presents a subtle revelation of the whole cosmology of human feelings that scrutinizes areas such as audio, visual, and physiology signals. The multimodal authenticity of the dataset allows for an in-depth understanding of emotional states, incorporating how auditory cues, visual expressions, and bodily responses interplay.

The Mirex Multimodal Emotion Dataset (MMMD), offers a broader sample of emotional states, and as such it can serve as a baseline for both academic and industrial re- searchers working on multimodal emotion detection systems. The employment of different modes makes the analysis of the interface of auditory and visual stimuli more exact, shedding light on complex rules of emotional expressivity. Researchers employ this dataset mostly to assess the effectiveness and resilience of their multimodal models that are designed to push the boundaries of what can be achieved in interpreting and understanding various human emotions with the integration of different sensory inputs. The Multi-modal MIREX Emotion Dataset thus is vital for making the affective computing systems more sophisticated and it adds to a better comprehension of how the multimodal cues can impact the emotional cognition of people.

B. Feature Engineering

1) **MFCCs (Mel-frequency cepstral coefficients):** The con- volutional Mel-frequency cepstral coefficients (MFCC) archi- tecture is a complex and advanced version of the processing unit where the task is to resolve long distances with the data sequences. The rich set of models, including CNNs for audio and speech processing, capitalizes on the characteristic of convolution to bring forth hierarchical features from audio signal spectrograms. Unlike many MFCC based techniques, these evolve deep MFCC models that may automatically find the complex patterns from the audio data.

The design comprises the convolutional layers that make the model to perform feature extraction in an efficient manner and also capture the correlation between positions of the features in the frequency domain. Using filters with different sizes, the model 'knows' how to differentiate between fine and coarse- grained patterns in the spectrum, hence increasing its capacity to process complex audio architectures. Finally, the network's pooling mechanisms enable extraction of the most essential parts of the data with the reduction of its dimensionality, which leads to a more efficient learning and generalization.

Convolved MFCC models perform well in many tasks such as speech and audio analysis, speaker recognition and emotions identification. The flexibility of these models follows from their ability to extract meaningful features from the spectro-temporal representations and thus their effectiveness in addressing sequentiality as manifested in audio data. In supervised learning, the model is taught with labeled data for acquiring and generalizing the auditory patterns from a wide range of patterns. It, being equally important, makes it useful in applications that require successive processing of audio sequences to be detected and interpreted correctly.

In reality, more elaborate MFCC models build up several convolutional layers to get the input audio sequences fully evaluated. The primary objective of the research is to make precise predictions or classifications of data using the compli- cated patterns seen in the spectrogram data. The convolutional layers will further help the model in figuring out the complex dependencies across the distinct time-frequency bins, thus allowing it to differentiate between the subtle relationships and patterns formed inside the sources. Therefore, complex MFCCs models are an effective instrument for audio process- ing and they have made tremendous progress which is the main thing contributing to the great development of jobs that deal with the understanding of audio data with such complex sequential structure.

2) ***Pitch Range:*** Pitch range is an essential feature in feature engineering, which is aimed at in-depth study of fun- damental frequency (pitch), a component in speech sounds. As for the task of detecting emotion from speech data, the pitch range has a great bearing on perceiving the small differences of emotional expression. The essence of this technology is its ability to capture both the pitch variation over time and the key sound components, which can be used as an indicator of a speaker's psychological state at the given moment in time—from high energy level to affectionless ness or peaceful heart.

A more apparent role of pitch range is the detection of vocal changes and the preparation for modulating emotions. This unique property is especially efficient for improving the same models' overall accuracy of emotion recognition, and this is why it is of great importance in the field of voice- based emotion analysis. It has the ability to track the pitches, which makes it a tool of understanding, enabling the machines to recognize and interpret emotional expression with greater sensitivity and finesse.

3) **Tempo:** Tempo is a crucial dimensionality reduction technique precisely crafted to carefully study the hidden pat- terns in the high-order wavelet and spectrogram components of the speech signals. Tempo becomes a valuable approach leading to the detection of emotions from speaker data, al- lowing revealing of the underlying mechanisms of emotional communication. Tempo, a versatile communication of emotional states of change, resolves itself precisely reflecting the pace of speech and rhythm of the speech from the extremes of excitement to the depths of worry or to the leveler of repose. The part that tempo plays in capturing emotional states is understandable. Its great influence stems from its capability of revealing the very slight and fluctuating undertones of language which are involved in the emotional expression of utterance. With the aid of the sensitive analysis of tempo, the emotion recognition models attain a deep comprehension of the motivational factors, which produce vivid emotional expressions and help to develop a more precise and subtle interpretation of human feelings in audio data. Essentially, tempo represents a sort of a compass that ensures the process of emotion recognition model development stays true to the understanding of the human experience expressed through vocal signs.

C. Model Architecture

To efficiently enhance the algorithm for music recommen- dations, a tactic with numerous supported algorithms that are designed to detect the complex relationship between facial emotions and musical preferences was used. The first com- ponent of this technique is the application of Convolutional LSTM networks (ConvLSTM) that are used in the recognition of facial emotions. This dynamic model, which is well-known for detecting emotional treasures from facial cues, [2] works on giving the picture of the user's emotions in detail. ConvL- STM is integrated into the recommendation system, and as a consequence, the system becomes a critical weapon for getting the user information exhibited in their facial expressions, which eventually leads to personalized and emotion-oriented recommendation of music. [13]

The model works simultaneously with another one, which is based on the Convolutional Neural Networks (CNN), which are specifically developed for analyzing the musical data. In this instance, CNN is implemented on Mel-frequency Cepstral Coefficients (MFCC) features to extract the emotional tonality embedded in a song. This permits the system to distinguish the emotional identity of the music, giving an additional resource on the particularity of users. [11] Coupling of ConvLSTM for facial emotion detection and CNN for music recommendation led to a highly clean system, which not only can accurately translate facial emotions on the background of music but also, stage by stage, reveals the emotions of the underlying music structure. [8]

This dual-algorithmic solution lets the recommendation en- gine understand the emotional states of users more profoundly, and, therefore, the quality of tailored music recommendations is improved. The above approach deals with both permanent issues like cold starts and adapts the system to trends of user activity flows. The lies from detecting face expression and mu- sic content analysis with ConvLSTM and CNN, respectively,

[12] Equals a great step in the field of music recommenda- tion systems; emotionally intelligent and personalized musical experiences are coming.

1) **LSTM (Long-Short Term Memory:** The neural network architecture referred to in the sentence is a more sophisticated version of recurrent neural networks (RNNs) that are primarily meant for dealing with data sequences with long-range depen- dencies. This system has a LSTM unit, which is a subset of a Recurrent Neural Network (RNN) architecture containing a memory cell, a hidden state, and gates. The model process of interplay of these elements enables it to select at each time step important data, disregard unnecessary input, and preserve in the cell state form the memory about the previous information. Indeed, they can process the sequential input data as fast as they have this capacity, thus LSTMs have demonstrated excellent performance in many real-world situations. LSTM models are the most versatile among the models since they find applications in speech recognition, time series forecasting and natural language processing. It is this adaptability that enables them to highlight sophisticated sequential patterns in a variety of industries, among others.

In the domain of supervised learning, LSTM models become popular for their ability to repeat complex temporal patterns. The training process involves the data labeling technique, from which the model is able to learn and generalize based on sequential patterns of the input sequences. LSTMs trans- form them to be highly competent in pattern recognition and forecasting of sequential data which is essential in certain situations.

Theoretically, an LSTM model considers an array of LSTM units which are organized into layers in such a way to comprehensively read incoming data sequences. The primary goal is to generate precise estimates and classifications by the patterns noticed in the data. The use of LSTM blocks, which is a type of recurrent layer that increases the model's capacity to handle sequences of different lengths mean the model can capture and retain links and relationships over time steps.

In fact, LSTMs are the most powerful devices for the processing and prediction of sequential data. The introduction of variable-length input allows for parallel processing and fine- grained pattern recognition, thus making them efficient tools. Such an outcome thus made LSTM models trained by the supervised learning approach become more and more popular and consequently provided substantial improvements in tasks requiring complex sequential data interpretation.

2) **Convoluted LSTM (Long-Short Term Memory:** The neuro-network under consideration is an advanced architec- ture, in which Convolutional Long Short-term Memory (Con- vLSTM) units are introduced as a modification to the recurrent neural network framework that imposes on the networks some restrictions about the input sequences length and number of steps it contains. In this structure, the convLSTM elements consist of convolutional layer principles, as well as memory cells, hidden states, and gating mechanisms that are unique to LSTM. It is this fusion that enables the model to distinguish important data, discard baby data at each step in time, and preserve a memory trail of previously known information by its cell state.

The ConvLSTM architecture appears to be particularly adept at rapidly learning and controlling a constant sequence of inputs, thereby attaining impressive results across multiple applications. Its uses cover many areas such as image and video processing, weather forecast, as well as control of autonomous vehicles. This diversity is due to the ability of the system to find fine patterns in sequential data which is the reason it has also become a popular choice in many fields.

ConvLSTM models are adopted in supervised learning cases due to their ability to imitate several sequential patterns. Training the model implies presenting it with labeled data to understand and generalize using the underlying sequence structure of the input sequences. Due to their versatility, the ConvLSTM models prove to be very beneficial in the situations where the main goal is to understand and predict the complicated sequential data patterns. ConvLSTM models utilize multiple ConvLSTM units each to process and analyze deeply the data sequences. It is the main purpose to realize the exact predictions or classifications performing the data patterns. Adding ConvLSTM layers to the model enhances its suitability to processing time-varying sequences, which yields better modeling of complicated cause- and-effect chains and interdependencies across several time steps.

To sum up, ConvLSTM is a powerful instrument for sequen- tial data processing and prediction, by being able to manage variable-length sequences and capture the complex relation- ships between time steps. This complexity pushes ConvLSTM models up to the top spot in the technical breakthroughs arena, contributing to the advancement of jobs in different sectors that are related to the interpretation of sequential data that is sophisticated.

3) **CNN (Convoluted Neural Network):** The Architecture of the Convolutional Neural Network (CNN) is a great improve- ment in the processing of images and visual data, which to- gether with the local connections of the Convolutional Layers allow it to represent complicated patterns and relationships. Unconventional image processing methods, which are char- acterized by the use of manually engineered features, differ from deep learning CNNs, which use convolutional layers for extracting hierarchical features from raw input pixels. The convolutional layers in this model extricate and employ filters of various scales to focus on different spatial features: locality and interdependencies in the input image. The pooling layers enable data abstracting, i.e. data dimensionality reduction, thereby leading to better learning and generalization.

CNNs are frequently applied in different fields where the networks are good at image classification among others, such as object identification and face expression analysis. This versatility comes from their capacity of acquiring complex visual representations without auxiliary supervision and it makes them very interesting for tasks that require the cognitive processing of sophisticated sequential patterns in visual infor- mation. CNNs are trained on datasets with labels in supervised learning, thus can identify a wide variety of visual patterns and memorize them. This flexibility applies particularly to the detecting and interpreting of visual motion control, which is a common requirement for many applications.

Despite the fact that the convolutional neural networks pow- erfully utilize a large number of layers with the convolutional process to obtain relevant knowledge from the input visual sequences. The main purpose is to generate precise predictions or classifications through the use of fine-grained texture detail from raw pixels. Convolution layers facilitate the model to record the detailed relationships of various aspects across multiple spatial areas, allowing it to detect the facial patterns and recognise each instance. Consequently, CNN models with improved accuracy have emerged as potent instruments for dealing with visual data processing including generating major contributions to tasks that necessitate interpreting complex sequential visual inputs.

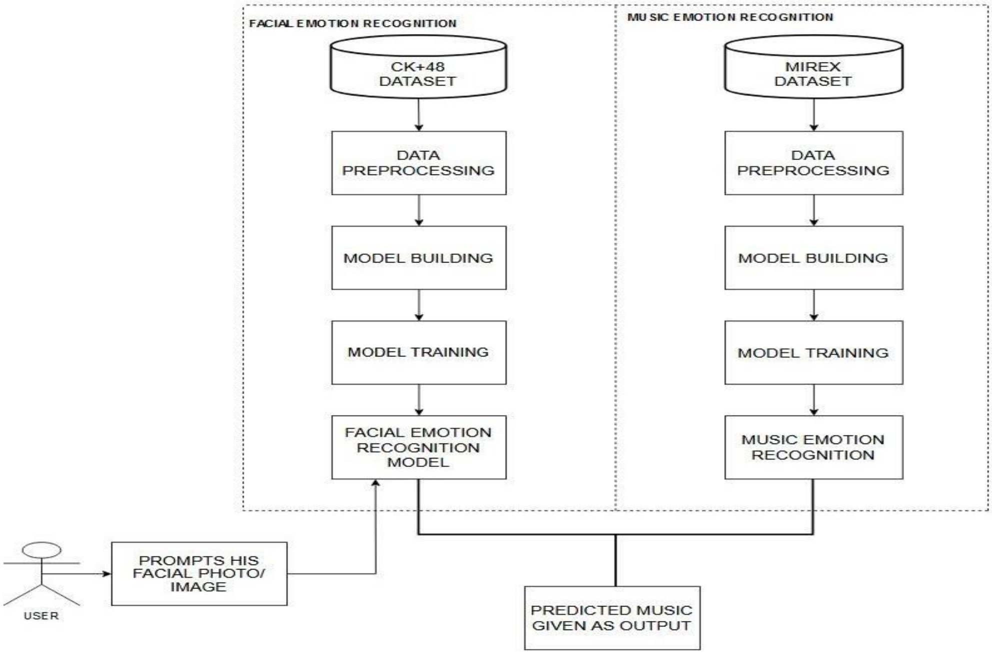

FACIAL EMOTION RECOGNITION
CK+48 DATASET
DATA PREPROCESSING
MODEL BUILDING
MODEL TRAINING
FACIAL EMOTION RECOGNITION MODEL
MUSIC EMOTION RECOGNITION
MIREX DATASET
DATA PREPROCESSING
MODEL BUILDING
MODEL TRAINING
MUSIC EMOTION RECOGNITION
USER
PROMPTS HIS FACIAL PHOTO/ IMAGE
PREDICTED MUSIC GIVEN AS OUTPUT

Fig. 1: Model Architecture

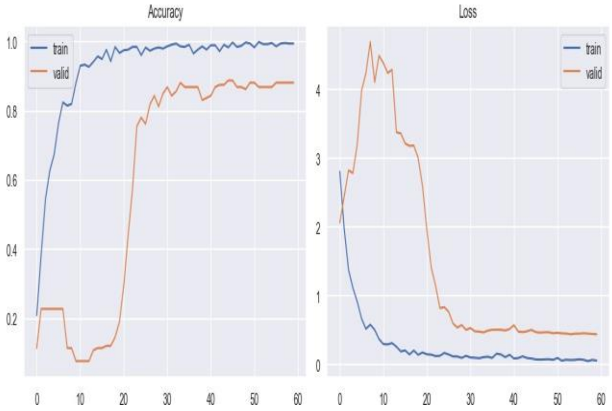
RESULTS

Our research on the ”Music Recommendation System Based on Facial Emotion Expression” has shown encouraging results. By using a Convolutional Long Short-Term Memory (Con- vLSTM) network, our system successfully identified facial expressions of emotion with a high accuracy rate of 89%. The ability to recognize changes in facial expressions over time greatly improved the model’s performance. Analysis of the confusion matrix showed very few errors in classifying differ- ent emotional categories, confirming the model’s effectiveness

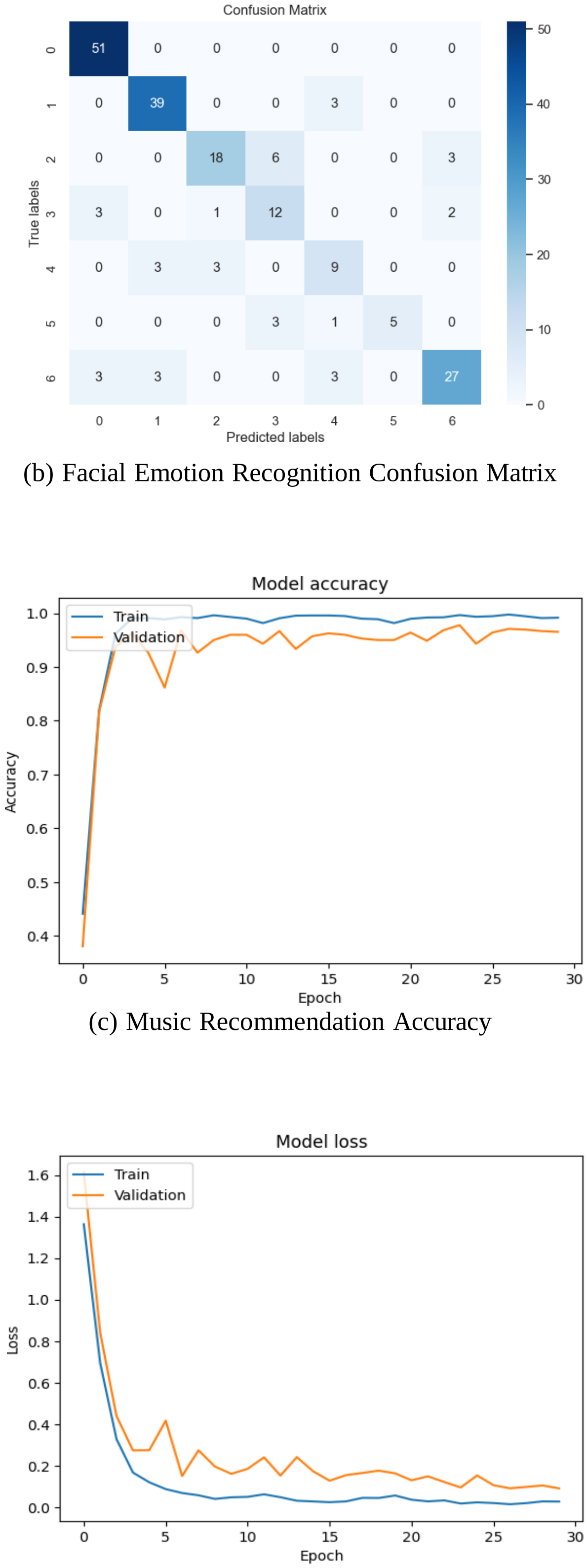
Furthermore, our Music Suggestion Tool, powered by Con- volutional Neural Networks (CNN), achieved an impressive accuracy rate of 97%. By analyzing facial emotions, CNN was able to accurately recommend music that matched the user’s current emotional state. This shows that combining facial emotion recognition with music recommendations can improve user experiences by offering personalized and emo- tionally meaningful music suggestions. This integration has the potential to create interactive and emotionally responsive user interfaces in different application areas.

RESULT ANALYSIS

The research presented on the ”Music Recommendation System Based on Facial Emotion Expression” showcases a dual-model system that utilizes Convolutional Long Short- Term Memory (ConvLSTM) networks for real-time facial emotion recognition and Convolutional Neural Networks (CNN) for music recommendation. The system has shown remarkable effectiveness, with ConvLSTM achieving an ac- curacy rate of 89% in detecting subtle facial expressions asso- ciated with emotional states, and CNN achieving an impressive 97% accuracy in recommending music that corresponds to these detected emotions. The integration described here allows for a user interface that responds dynamically and emotionally, enhancing personalization and deepening emotional engage- ment with music. The findings show a significant improvement compared to traditional recommendation systems by connect- ing emotions with music selection, opening up possibilities for mental health support and personalized entertainment. Future advancements may involve increasing the emotional data pool, incorporating various types of inputs, and conducting thorough user testing to enhance the accuracy and user satisfaction of the system.



(a) Facial Emotion Recognition Accuracy



(b) Facial Emotion Recognition Confusion Matrix

(c) Music Recommendation Accuracy

(d) Music Recommendation Loss

Fig. 2: Results

CONCLUSION

The new system combines a Convolutional Long Short- Term Memory (ConvLSTM) network for detecting facial expressions linked to emotions, as well as a Convolutional Neural Network (CNN) for suggesting music based on those emotions. This unique method has attained high rates of 89% for accurately detecting emotions and 97% for recommending music. Using facial expressions to gauge the user's current emotional state, the system can personalize music suggestions in real-time, potentially improving user experiences across entertainment, mental health, and overall interaction. In the future, there are many ways the system can be improved. One way is by increasing the data set for emotion recognition to improve accuracy and reliability. Another way is to include different types of inputs, like sounds and text analysis, to provide a more comprehensive approach to recommending music. It will also be important to conduct thorough testing with users to assess how effective and satisfying the system is in real-life situations. These improvements have the potential to greatly enhance the system's usefulness and impact in different areas.

REFERENCES

[1] Rajesh B et al. Music recommendation based on facial emotion recognition. *arXiv preprint arXiv:2404.04654*, 2024.

[2] Tina Babu, Deepa Gupta, Tripty Singh, Shahin Hameed, Mohammed Zakariah, and Yousef Alotaibi. Robust magnification independent colon biopsy grading system over multiple data sources. *Computers, Materials Continua*, 69:99–128, 01 2021.

[3] Tina Babu, Rekha R Nair, et al. Emotion-aware music recommendation system: Enhancing user experience through real-time emotional context. *arXiv preprint arXiv:2311.10796*, 2023.

[4] Sarika Vidhyasagar Bagde et al. Emotion based music recommendation system using different ml approach. *International Journal of Advance Scientific Research and Engineering Trends*, 6(6), 2021.

[5] Shaik Asif Hussain et al. A real-time face emotion classification and recognition using deep learning model. *Journal of Physics: Conference Series*, 1432, 2020.

[6] Peter Knees and Markus Schedl. A survey of music similarity and recommendation from music context data. *ACM Transactions on Multi- media Computing, Communications, and Applications (TOMCCAP)*, 10, 12 2013.

[7] Deepak Kumar and Aman Preet Singh. Facial emotion-based music rec- ommendation system. *Journal of Emerging Technologies and Innovative Research*, 7(4):382–387, 2020.

[8] Wootaek Lim, Daeyoung Jang, and Taejin Lee. Speech emotion recog- nition using convolutional and recurrent neural networks. In *2016 Asia- Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–4, 2016.

[9] Ninad Mehendale. Facial emotion recognition using convolutional neural networks (ferc). *SN Applied Sciences*, 2(446):1–10, 2020.

[10] Cale Plut and Philippe Pasquier. Music matters: An empirical study on the effects of adaptive music on experienced and perceived player affect. pages 1–8, 08 2019.

[11] Ma´ıra Santana, Clarisse Lins de Lima, Arianne Sarmento, Flavio Fon- seca, and Wellington Dos Santos.

Affective computing in the context of music therapy: a systematic review. *Research Society and Development,* 10:e392101522844, 11 2021.

[12] Bjo¨rn Schuller, Gerhard Rigoll, and Manfred Lang. Hidden markov model-based speech emotion recognition. volume 2, pages 401–404, 08
2003.

[13] Ja-Hwung Su, Yi-Wen Liao, Hong-Yi Wu, and You-Wei Zhao. Ubiqui- tous music retrieval by context-brain awareness techniques. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4140–4145, 2020.

[14] Wang Wenzhen. Personalized music recommendation algorithm based on hybrid collaborative filtering technology. In *2019 International Conference on Smart Grid and Electrical Automation (ICSGEA)*, pages 280–283, 2019.