Music Recommendation Filter Based on Emotions Extracted from Facial Expression and Speech Audio

Group 3: HUANG HAI-HSIN, TSAO KAI-TING, YINGGE XU, KANG YIFAN, MA KE
GitHub Link: github.com/BT_5153/Group_3/final_project

1. Introduction

Music is universally recognized as an effective way for humans to express emotion and regulate emotional states. Studies show that people can evoke 13 distinct emotions from listening to music. Therefore, besides daily refreshment, music is also widely used in professional areas such as music therapy and music education. Most of the existing music recommendation engines are based on collaborative preference or music content, without considering human emotions. However, we think that what people need is not only music, but music fitted for the current emotion. We aim to create an emotion-based music recommendation model layer to better empower the existing system, which could bring lots of value.

1.1 Use Case

The use cases could be summarized into two types, the passive use case, and the active use case.

- Passive Use Case:

These are cases where we collect the users' emotional data for analysis, mostly in public places where privacy is not a big concern. Some of the passive cases digest the information, which in turn protects users, bringing needed help in their work. For example, this layer can be embedded in the safe driving system to detect changes in the driver's expressions and voices in real-time, to avoid fatigue driving or conflicts. Another example is that some cinemas automatically monitor and collect user expressions to get feedback on user experience.

- Active Use Case:

In most causal use cases, there is a huge difficulty in collecting emotional data since most of the users do not want to share the data. Therefore, asking for privacy permission is a must. Users may be willing to share this data when looking for companions. For example, considering the old or the sick who need more delicate emotional care, and people who are eager to have more interactions when quarantined during the Covid-19 pandemic.

1.2 Business Value

From the users' perspective, it provides a sense of compassion and navigates listeners to a more positive emotional state, improving the user experience and gaining user satisfaction.

From the business side, high satisfaction means higher customer stickiness, as well as easier acquisition. This emotion-based layer could be embedded in various products, no matter software or hardware. For example, we already see many AI assistants, like Amazon Alexa, Google assistant, Xiaodu, Xiaoai, etc. One can interact with it in apps, also smart speakers, and home automation. During this implementation process, with the model trained and the data gathered, the business side could learn more about the user's behavior and psychology. Finally, the techniques, the products (both software and hardware), and various use cases together help the business to form its smart industrial ecology. Different from the previous ecology, this integration is more emotion-based, generating a more user-friendly environment.

2. Facial Expression Recognition

2.1 Data Description

Our facial expression dataset is Facial Expression Recognition Challenge Dataset[1] from Kaggle. This dataset has a uniform size for all images, 48x48 pixels. Pixel values of an image are extracted and stored in a csv file. Due to resources and time constraints, we will use the pre-processed csv file to conduct a descriptive analysis.

Each emotion is imbalanced in this dataset, as we show in ***Figure 1***. There is not enough 'disgust' emotion data, so we decide only to use the other 6 emotions,

[1] Facial Expression Recognition Challenge Dataset from Kaggle, https://www.kaggle.com/debanga/facial-expression-recognition-challenge

including anger, fear, happiness, sadness, surprise, and neutral, in our analysis.

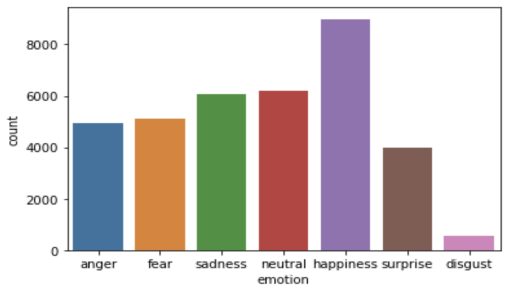

Figure 1: Number of Images by Emotion

Table 1: Facial Data Description

Feature	Type	Description
Emotion	Integer	Six categories: 0 = anger, 2 = fear, 3 = happiness, 4 = sadness, 5 = surprise, 6 = neutral
Usage	String	28,273 for training, 3,534 for validation, 3,533 for test
Pixels	String	Each image has 2,304 (48x48) pixels

2.2 Data Augmentation

To enhance the accuracy of image classification, we use the ImageDataGenerator method to augment data, including rotation, shifting width, shifting height, shearing, zooming, flipping, etc.

2.3 Modeling

This is a classification problem and the most promising machine learning tool for image recognition is Convolutional Neural Networks (CNNs)[2]. ***Figure 2*** is the sketch of image recognition problems using a CNN, consisting of image inputs, convolutional layers, fully connected layers, and output predictions.

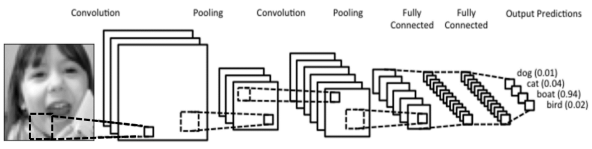

Figure 2. Sketch of Convolutional Neural Networks (CNNs)

Normal CNN generally has two or three layers but deep CNN will have multiple hidden layers, usually more than 5, which are used to extract more features and increase the accuracy of the prediction. There are two kinds of deep CNN, one is increasing the number of hidden layers or increasing the number of nodes in the hidden layer. In this case, we use the CNN model (***Table 2***) as a baseline model and further explore the deep CNN structure (***Table 3***).

Tabel 2. Hyperparameter Selection of CNN model

Hyperparameter/ Structure	Description
Structure	4 convolutional layers, 4 batch normalization layers, 2 max-pooling layers, 2 dropout layers, 1 flatten layer, 2 dense layers
Kernel Size	3
Pool Size	2
Activation Function	relu
Dropout Rate	0.25 for first dropout layer, 0.5 for second dropout layer
Batch Size	32

Table 3. Hyperparameter Selection of DCNN model

Hyperparameter/ Structure	Description
Structure	10 convolutional layers, 11 batch normalization layers, 4 max-pooling layers, 5 dropout layers, 1 flatten layer, 2 dense layers
Kernel Size	5 for first two convolutional layers, 3 for other convolution layers
Pool Size	2
Activation Function	elu
Dropout Rate	0.4 / 0.4 / 0.4 / 0.5 / 0.6
Batch Size	32

To avoid the model overfit the training dataset and have poor performance on the test dataset, we set

[2] Using Convolutional Neural Networks for Image Recognition, By Samer Hijazi, Rishi Kumar, and Chris Rowen, IP Group, Cadence, https://ip.cadence.com/uploads/901/cnn_wp-pdf

"early stopping". When the performance on a validation dataset starts to degrade, the model stops training at that point.

Also, we set the "ReduceLROnPlateau". When the performance on the validation dataset has stopped improving, the model would reduce the learning rate.

We summarize the performance of the CNN and deep CNN (DCNN) model in ***Table 4***. DCNN model gives us a higher classification accuracy both in the training set and the validation set.

Table 4: CNN & DCNN Performance

Model	Training Accuracy	Validation Accuracy
CNN	0.5854	0.6186
DCNN	0.7701	0.7071

Therefore, we select the DCNN model as the final one for the application. The training process is shown in Figure 3 and Figure 4. After 78 epochs of training, the performance becomes stable. The model stops at 89 epochs of training and restores model weights from the end of the best epoch.

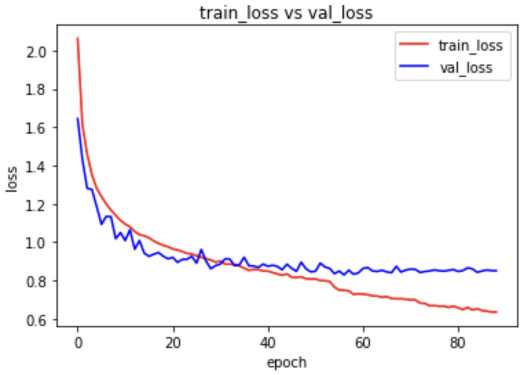

Figure 3. DCNN Loss of Each Epoch

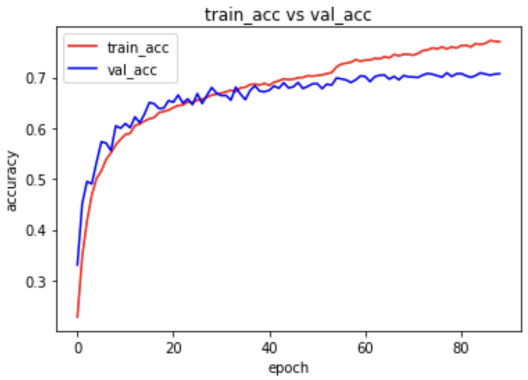

Figure 4. DCNN Accuracy of Each Epoch

2.4 Model Result

Table 5. DCNN model result

Emotions	Precision	Recall	F1-score
anger	0.72	0.58	0.64
fear	0.45	0.64	0.53
happiness	0.86	0.89	0.88
sadness	0.57	0.62	0.59
surprise	0.81	0.79	0.80
neutral	0.68	0.58	0.62

The final accuracy score of the facial expression recognition model is 0.69 and the final loss of this model is 0.91. We used this model to further predict the probability for each emotion, giving the final emotion category along with the probability level. The probability level was later plugged into the recommendation system as input.

3. Audio Emotion Recognition

3.1 Data Description

For the audio-emotion dataset[3], we use the RAVDESS Emotional speech audio dataset from Kaggle. This dataset contains 1440 speech audio files collected from 24 actors (12 males and 12 females). Each actor will record 60 audio trials with two lexically-matched statements in a neutral North American accent. The audio files have already been labeled with some features as below:

- Modality: 01 = full-AV, 02 = video-only, 03 = audio-only. All files in this dataset are audio-only files.
- Vocal channel: 01 = speech, 02 = song. All files in this dataset are speeches.
- Emotion: 01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fear, 07 = disgust, 08 = surprised. We will take the universal six emotions, including angry, fear, happy, sad, surprise, and neutral.
- Emotional intensity: 01 = normal, 02 = strong. Note that there is no strong intensity for the 'neutral' emotion.

[3] RAVDESS Emotional Speech Audio Dataset from Kaggle, https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio

- Repetition: 01 = 1st repetition, 02 = 2nd repetition.
- Actor: 01 to 24. ; Gender: male and female.

There are 192 files in each of the emotions in the dataset, except the neutral emotion containing only 96 files, as we show in ***Figure 5***.

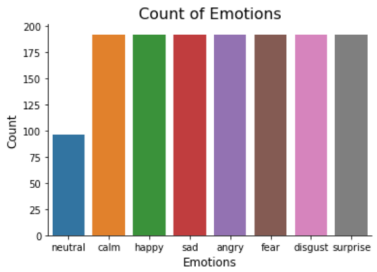

Figure 5 Audio Dataset Description

3.2 Data Augmentation

After the previous steps, we keep 1056 files, which is not enough for us to perform a deep learning analysis. Therefore, we create new synthetic data samples by adding small perturbations to our initial training set.

In this case, we try to add random noise, slow down the speech file by a rate of 0.8, change the pitch, and shift the time. After these steps, we get 3168 files in total. Besides, the augmentation also makes our model invariant to those perturbations and enhances its ability to generalize.

3.3 Audio-Features Extraction

Audio features[4] such as pitch, intensity, spectral energy distribution, average zero-crossing density (ZCD), jitter, and MFCC have been discovered to be useful in emotion recognition. However, using only one audio feature is inefficient to train emotion-recognition models. The trend of audio-features extraction is to combine several complementary audio features. Despite feature selection and engineering, the model accuracy rate is also found to be correlated to the number of emotions to be classified. Many models with higher accuracy rates classify only 2-3 emotion classes, indicating a tradeoff between model granularity and accuracy rate.

Table 6. Audio Feature Description

Audio Feature	Description	Formulation
Zero-crossing rate (ZCR)	The weighted average of the number of times the speech signal changes sign within a particular time window.	$\operatorname{sgn}\{x\} = \begin{cases} 1 & \text{if } x(n) \geqslant 0 \\ -1 & \text{if } x(n) < 0 \end{cases}$ $w\{n\} = \begin{cases} \frac{1}{2N} & \text{if } 0 \leqslant n \leqslant N-1 \\ -1 & \text{if otherwise} \end{cases}$
Chroma Stft	Performs short-time fourier transform of an audio input and maps each STFT bin to chroma	$w(n) = \sin\left(\frac{\pi}{N}(n+0.5)\right)$ $n = 0, \ldots, N-1.$
Mel-Frequency Cepstral Coefficients (MFCC)	Cepstral coefficients derived from a mel-scale frequency filter-ban	$w(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{N}\right), \quad 0 \leqslant n \leqslant N$ $Y(n) = X(n) \times w(n)$ $Mel(n) = 2595 \times \log_{10}\left[1 + \left(\frac{n}{700}\right)\right]$
Root mean square value	Measures the average loudness of an audio track within a time window	$\sqrt{\frac{1}{T2-T1}\int_{T1}^{T2}[f(t)^2 dt]}$
MelSpectogram	A spectrogram where the frequencies are converted to the mel scale.	$m = 2595\log_{10}\left(1 + \frac{f}{700}\right)$

Mean values for pitch features are shown in ***Figure 5*** and averaged pitch signals for each emotion on the Hamming window are shown in ***Figure 6***. The below graphs display great discrimination between the six emotions.

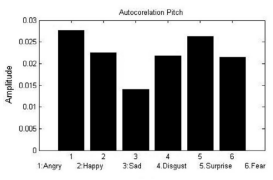

Figure 5. Mean Values for Pitch

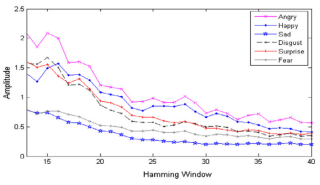

Figure 6. Hamming Window

[4] A new approach of audio emotion recognition, By Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, Li Wern Chew
A new approach of audio emotion recognition - ScienceDirect

3.4 Modeling

We first take some simple baseline models, like the regression models, SVM, etc. However, these models give a bad performance, whose accuracy is slightly higher than 50%. Considering the complexity of the audio dataset, we finally decided to train a Convolutional Neural Network model. We used the GridSearch method to choose the best hyperparameter and the network structure could be summarized below.

Table 7. Hyperparameter Selection

Hyperparameter/ Structure	Description
Structure	Four convolutional layers, four max-pooling layers, two dropout layers, one flatten layer, two dense layers
Kernel Size	8
Pool Size	4
Strides	1 for convolutional layer, 4 for pooling layer
Activation Function	"tanh"
Dropout Rate	0.1
Batch Size	80

After 40 epochs of training, the loss and the accuracy both become stable. The training process is summarized in ***Figure 7*** and ***Figure 8***.

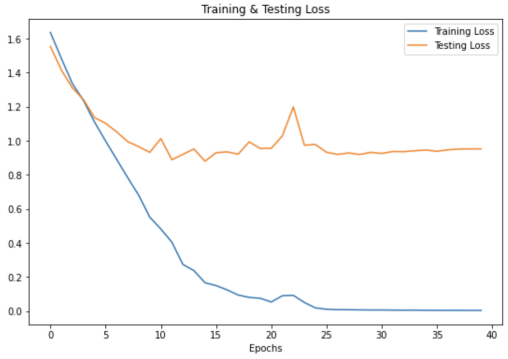

Figure 7. Loss of Each Epoch

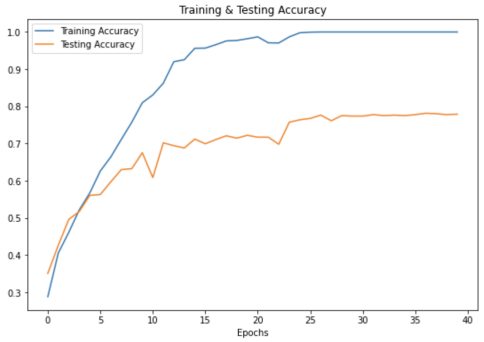

Figure 8. Accuracy of Each Epoch

3.5 Model Result

Table 8. Model Result

Emotions	Precision	Recall	F1-score
angry	0.91	0.80	0.85
fear	0.78	0.84	0.81
happy	0.73	0.68	0.71
neutral	0.73	0.73	0.73
sad	0.73	0.82	0.77
surprise	0.79	0.78	0.79

The final accuracy score of the audio-emotion recognition model is 0.78, and we used this model to further predict the probability for each emotion. The probability level was later plugged into the recommendation system as input.

4. Recommendation System Building

4.1 Matching Strategy: Valence-Arousal Space

The goal of emotion-based music recommendation is to guide the user to a more positive emotional state.

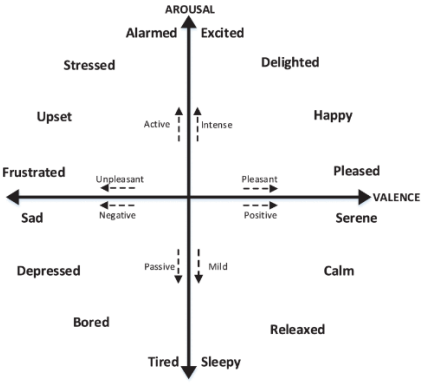

Figure 9: Valence-Arousal of Emotions

However, Mou's research suggests recommending a piece of music that has a similar Valence-Arousal value to the user's self-reported emotional state as the first piece of music that the user listens to. This expression of similar emotion can give the user a feeling of compassion. Especially when a user is experiencing a depressed mood, the attempt to positively influence the user by a piece of music showing compassion may be more effective than directly playing an exciting or joyful one. In the common approach, to simplify the recommender for demonstration, we directly match emotions by similarity of emotions from music and user's current status. Besides, there may be other scenes with a more specific demand. For example, music therapy follows progressive emotion arousal and vehicle-mounted application scenarios focus more on driving safety where fatigue driving can be detected and responded to with alarm or some uplifting music.

4.2 Output Emotion Integration

To integrate emotion results generated from facial and audio data, we execute the following steps:

- For one moment, we capture the user's current facial expression images and audio to train two models separately and output probabilities for six emotion categories.
- Mapping the probability vector of six categories into Valence-Arousal space (by ***Table 9***)
- Combined the output emotion label as weighted average by **f1** score for a certain emotion from two models.

$$(x,y) = p_f^T \cdot (x_f, y_f) \times \frac{F1_f}{F1_f+F1_a} + p_a^T \cdot (x_a, y_a) \times \frac{F1_a}{F1_f+F1_a}$$

(x,y): The weighted score of valence-arousal combining two models.

(x,y): The valence-arousal matrix of the six emotions.

p: probability vectors of the 6 emotion categories.

F1: The model f1-score for a specific emotion.

The suffix ***f*** represents results from the facial-emotion model and the suffix ***a*** represents results from the audio-emotion model.

We are doing this weighted average by assuming that in terms of one label, the higher the F1 score, the more confident a model is. The prediction probabilities represent the intensity of different emotions.

Table 9. Mapping Emotions into Valence-arousal[5]

Emotions	Valence	Arousal
angry	2.81	5.03
fear	2.59	4.33
happy	5.37	4.99
neutral	5.93	2.89
sad	5.05	2.55
surprise	3.10	4.46

This 2D matrix is obtained from ***Soundtracks*** dataset, where intensity ratings for more than 7 emotions are collected from 24 professional musicians after they listen to over 200 soundtracks from famous movies. The ratings dimensions include valence and arousal (range from 0~9) as well. And the targets of each piece of music cover our six emotion categories. So, by calculating the overall average valence-arousal for each emotion, we built this matrix.

4.3 Music Recommender Design

Chill	Commute	Energy Boosters	Feel Good
Party	Romance	Sleep	Workout

***YouTube Music** recommend by mood & moments*

Through facial expression recognition and audio emotion recognition systems, information about the user's current emotion will be fed to any collaborative or content-based music recommender as supplementary features. We expect it can boost conventional recommender by improving recommended precision.

[5] Valence-arousal scores for emotions are gathered from the dataset, ***Soundtracks***, Eerola, T. & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. Psychology of Music, 39(1), 18-49. https://doi.org/10.1177/0305735610362821
Detailed page of the dataset: https://osf.io/4wzc9/

We assume two ways for music recommendation engines to collaborate with user emotion data.

(1) Algorithm Boost: Feature matrix integration

The proposed system is an interactive mood-based songs recommendation system that considers the current emotion of the user along with vital factors related to songs and user preferences for these factors while recommending songs. When the user gives feedback (thumb, score…), labels are generated. Recommender then retrains the model to learn the user preference pattern with moods. The recommender performs better as users interact more.

(2) Emotion Time: Directly mapping music by the current mood

The second way is directly mapping music with the user's current mood through a matching algorithm. As mentioned above, emotions can be deconstructed into 2 dimensions (Arousal and Valence). All moods will be mapped with labeled music in the dataset. The mapping process required the involvement of musicians and volunteers. To simplify this complex process, our recommender maps music in a similar mood to resonate with users.

4.4 Music Emotion Recognition Dataset

Algorithm Boost is built on the base of a comprehensive music recommender which is not the focus of our research. In order to deploy and validate our design, a demonstration of the **Emotion Time** music recommendation system was established. We pick **SoundTracks** as a sample music dataset as it contains fruitful soundtracks from famous movies and all are labeled by valence-arousal scores.

This large study established a set of 110 film music excerpts, half were moderately and highly representative examples of five discrete emotions (anger, fear, sadness, happiness, and tenderness), and the other half were moderate and high examples of the six extremes of three bipolar dimensions (valence, energy arousal, and tension arousal). These excerpts were rated in a listening experiment by 116 non-musicians and take average ratings for Valence, Energy, Tension, Anger, Fear, Happy, Sad, Tender. Finally, the emotion with the highest ratings is assigned to this piece of music as the target.

4.5 A Demo for Emotion Time Recommender

Because of dataset constraints, we are not able to get streaming facial expression images and audio for one user. Thus, we simplify the process and build a demo program to demonstrate how the Emotion Time recommender works.

In the demo, we pick one result probability vector from the facial recognition model. After mapping it into valence-arousal space, we calculate its cosine similarity with every piece of music in the dataset and filter the top 3 pieces of music with the largest similarity to recommend[6]. From the testing results of the demo program, most users are recommended by music in a similar mood.

```
User's valence-rousal score:
Valence    5.364372
Arousal    4.988493
Name: 3, dtype: object

Predictied emotion: happiness

Top3 Recommend music:
#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-
   number  valence  energy  similarity   Emotion           Album name  Track  \
0      27     4.60    4.20    0.999958     Happy        Shallow Grave      8
1     153     3.83    3.67    0.999888  Surprise          Naked Lunch     14
2      22     5.60    5.00    0.999795     Happy  Shakespeare in Love     21

       Min:Sec
0  00:00-00:15
1  01:02-01:17
2  00:03-00:21
#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-#-
```

Figure 10. Top3 Recommended music for No.4 user in facial dataset

5. Limitation and Future Improvements

The general analysis framework has been constructed up to now. However, if we want to put this framework into industrial realization, there are still some limitations calling for future improvements.

5.1 Datasets

We have concluded three biggest problems in our framework as below: first and foremost, our facial dataset and audio dataset are collected from different sources, and hence they are also from different people. For this reason, we are not able to merge these two datasets together to apply the formula presented in section 4.2. However, this is not a big issue for real business use cases because it is possible for customers to agree on providing facial and audio data authorization at the same time.

Secondly, our datasets are not comprehensive enough, which may contain biases. In the current datasets, the face images are basically taken from the front side, and the audio dataset simply includes audio records from speeches. However, it is not possible, nor is it appropriate to ask users to take photos at this exact angle or speak in an exact tone. Thus, in the training stage, the data should include more possibilities for wider use cases. For example, facial images from all angles of a face can be trained to recognize facial emotion, and the model should also be able to tell

[6] See the full demo program here

emotions from a person with and without wearing glasses. Audio data used for model training purposes might include daily talks, which provide more noise than speech scenarios. The more input cases are included in the training and model construction part, the better product experience customers will have in the future application.

The last problem regarding data is the limitation of the target variables. Although we have 6 types of emotions, it will be more helpful if levels of emotional intensity, such as “happy – high” and “happy – moderate” instead of simply “happy", are included in the labels. If we can provide more accurately predicted emotion classes, there will likely be more benefits for real applications and usage.

5.2 Models

Secondly, our models have some limitations. Although the overall accuracies of the two models seem good, we wonder if they are high enough for business purposes. Moreover, as presented in sections 2.4 and 3.5, we can see that our facial and audio emotion recognition models have different predictive power for different emotions. This might provide some difficulties for customers under certain circumstances in usage. To tackle this limitation, other possible data augmentation methods and more advanced predictive models from papers could be applied. These new methods might be useful even if we use more data later to exclude possible training data selection biases.

5.3 Music Recommendation

Thirdly, some improvements could be made to the recommendation strategy. With the limited data resource, we can only generate the formula in section 4.2. Besides, we use cosine similarity to decide which pieces of music to recommend. It is reasonable since if the models predict that the user is angry now, this recommender will not play a happy song. However, if available in industries, some other mechanism could be designed based on insights from psychological research and user feedback from industry experiences.

Last but not least, we need to consider the variety of real-world use cases and make adjustments accordingly. If our recommendation layer is implemented into a car, it should not only play music to avoid fatigue, but also make some alerts if the driver loses attention. If this system is applied in an AI assistant designed for family usage, more features than simply emotion can be inputted to improve the enjoyment in a home circumstance. What we have provided in this project, considering these real cases, is a baseline framework with huge potential in real industries. This implies that future works are needed for more specific cases and considerations.

6. Conclusion

In this project, we have successfully built a framework for music recommendation with facial and audio emotion recognition.

For facial emotion detection, we explored the CNN and DCNN models, finally reaching an average accuracy of 69%. For audio emotions, we put much effort into data augmentation and feature extraction. Though the audio data is more complex, we reached an average accuracy of 78% with the CNN model. Recognizing emotions from facial expressions and audio speech is very useful in real-world cases.

Our project mainly focuses on music recommendations based on the emotion we have detected. We deep-dived into the professional area of psychology, especially the Valence-Arousal space. Depending on our research on various music databases and the matching strategies, we are able to build a demo program recommending music of similar mood to the current emotion.

Though we simplify some procedures, and there are improvements to be done, the emotion-based music recommendation layer can work smoothly, acting as a powerful framework. Furthermore, it is of good potential for modification for industrial specific needs. Especially for the emotion detection techniques, which could be widely applied to other products or other industries to improve product functions and customer satisfaction efficiently.

Reference

[1] L. Mou, J. Li, J. Li, F. Gao, R. Jain and B. Yin, "MemoMusic: A Personalized Music Recommendation Framework Based on Emotion and Memory," 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), 2021, pp. 341-347, doi: 10.1109/MIPR51284.2021.00064.

[2] F. Kuo, M. Chiang, M. Shan and S. Lee, "Emotion-based music recommendation by association discovery from film music", Proceedings of the 13th annual ACM international conference on Multimedia. Association for Computing Machinery, pp. 507-510, 2005.

[3] D. Ayata, Y. Yaslan and M. E. Kamasak, "Emotion Based Music Recommendation System Using Wearable Physiological Sensors," in IEEE Transactions

on Consumer Electronics, vol. 64, no. 2, pp. 196-203, May 2018, doi: 10.1109/TCE.2018.2844736.

[4] Song, Yading and Simon Dixon. “PREDICT THE EMOTIONAL RESPONSES OF PARTICIPANTS ?” (2015).

[5] F. H. Rachman, R. Samo and C. Fatichah, "Song Emotion Detection Based on Arousal-Valence from Audio and Lyrics Using Rule Based Method," 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), 2019, pp. 1-5, doi: 10.1109/ICICoS48119.2019.8982519.

[6] J. Kim, S. Lee, S. Kim and W. Y. Yoo, "Music mood classification model based on arousal-valence values," 13th International Conference on Advanced Communication Technology (ICACT2011), 2011, pp. 292-295.

[7] J. Bai et al., "Dimensional music emotion recognition by valence-arousal regression," 2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), 2016, pp. 42-49, doi: 10.1109/ICCI-CC.2016.7862063.

[8] A. Kolli, A. Fasih, F. A. Machot and K. Kyamakya, "Non-intrusive car driver's emotion recognition using thermal camera," Proceedings of the Joint INDS'11 & ISTET'11, 2011, pp. 1-5, doi: 10.1109/INDS.2011.6024802.

[9] Morrison D, Wang R, De Silva L C. Ensemble Methods for Spoken Emotion Recognition in Call - Centers [J]. Speech Communication, 2007, 49(2):98-112.

[10] Build Your First Mood-Based Music Recommendation System in Python

https://towardsdatascience.com/build-your-first-mood-based-music-recommendation-system-in-python-26a427308d96

[11] Tidke, S., Bhutkar, G., Shelke, D., Takale, S., Sadke, S. (2022). Mood-Based Song Recommendation System. In: , *et al.* Human Work Interaction Design. Artificial Intelligence and Designing for a Positive Work Experience in a Low Desire Society. HWID 2021. IFIP Advances in Information and Communication Technology, vol 609. Springer, Cham. https://doi.org/10.1007/978-3-031-02904-2_9

[12] A. Aljanaki, F. Wiering, R. C. Veltkamp. Studying emotion induced by music through a crowdsourcing game. Information Processing & Management, 2015.

[13] Deep Convolutional Neural Networks

https://www.sciencedirect.com/topics/computer-science/deep-convolutional-neural-networks

[14] Speech Emotion Recognition with CNN | Kaggle