APS360 PROJECT FINAL REPORT

Jonas Martins
Student# 1006869907
jonas.martins@mail.utoronto.ca

Chielotam Agbatekwe
Student# 1006988057
chielotam.agbatekwe@mail.utoronto.ca

Jennifer Sunny
Student# 1006998732
jennifer.sunny@mail.utoronto.ca

Ausef Yousof
Student# 1008470110
ausef.yousof@mail.utoronto.ca

ABSTRACT

This project analyzes the profound influence of music on human emotions and behaviors, aiming to utilize deep learning for music emotion analysis. By categorizing music into distinct moods, the initiative seeks to enable users to curate playlists that resonate with their emotional states, enhancing their listening experience. The research involves both traditional Machine Learning and Deep Learning methodologies to recognize and classify the emotions evoked by music. In our report, we discuss the intricacies of Music Emotion Recognition (MER), highlighting the significance of various musical aspects such as rhythm, harmony, tempo and pitch. We lay out a cohesive data processing method that extracts four feature spectrograms from raw music data, and model architecture consisting of a Convolutional Neural Network (CNN) followed by a Recurrent Neural Network (RNN). The project also emphasizes the ethical considerations, particulary concerning copyright issues, and outlines a comprehensive project plan to ensure efficient collaboration and timely completion. Our goal is to harness the power of deep learning to provide a refined understanding of music's emotional impact, bridging the gap between music, technology, and human emotion.

—-Total Pages: 9

1 INTRODUCTION

Music has a significant influence on human emotions and behaviours. Exposure to certain melodies, rhythms, and harmonies amongst other things can help alleviate bad moods and increase productivity. Music can also serve as a means of expression, a way to find community and understanding. The goal of this project is to employ music sentiment analysis to categorize music into genres or moods. The practical application of this is to grant a user the ability to curate playlists based on their desired emotional state in hopes of enriching their experience. Since music is not intuitively quantitative, there are several different features, such as tempo, amplitude and lyrics, that must be extracted from a track for analysis. The process of identifying feature importance (weights) and their relationships relative to each other to determine the resulting emotion can be extremely complex and hard to compute for traditional machine learning. Through the use of deep learning, these relationships can be easily captured with increased accuracy, scalability and potential to expand into further multi-dimensional analysis. The goal of our project is to create a model, EmotioNet, which successfully classifies a track in one of our four emotion classes 'Happy', 'Sad', 'Aggressive' and 'Calm' based on a given 45-second snippet of the track.

2 ILLUSTRATION

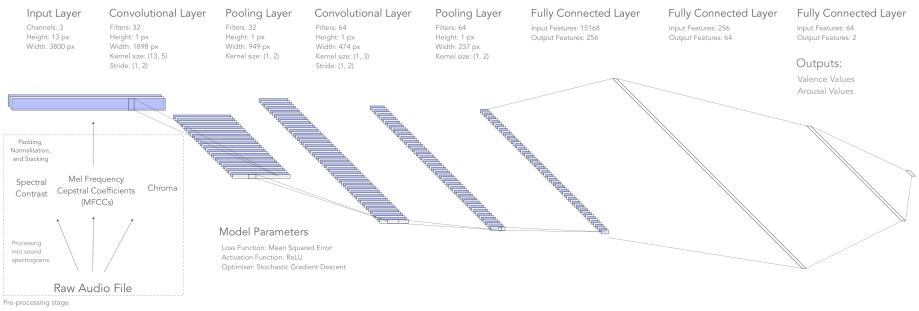

Figure 1: Visualisation of Model

3 BACKGROUND AND RELATED WORK

3.1 FEATURE EXTRACTION AND TRADITIONAL MACHINE LEARNING

"A survey of music emotion recognition" Han et al. (2022) explains how training a traditional machine learning model can classify Music Emotion Recognition (MER). This kind of model requires two parts. First for specific features to be extracted and then to be analyzed by an algorithm. Since music is not intuitively quantifiable, researchers have selected certain aspects of music which can be extracted. Features of the track's audio feature (waves, timbre), symbolic feature (pitch, interval, duration). An emotion model must also be chosen so that the sampled music can be sorted into clear categories. The most commonly used by deep learning models to categorize perceived emotions evoked by music is the Russell's Circumplex Model. Emotion is determined by valence (positive/negative emotion) and arousal (passive and activated emotion evoked). Features are extracted using preprocessing methods such as framing, windowing, spectrogram extraction, main track extraction and preprocessing tools such as Pysound, MATLAB, or with python packages like Librosa, pretty_music, music21 The preprocessing results in MFCC, spectrogram, key, BPM, melody and other musical features which are then fed into machine learning levels. They are then fed into machine learning models, and will result in different types of classifications, depending on the model. One type is Song Level Categorical MER whose classification model is support vector machines (SVM) and uses k-nearest neighbors, decision tree, random forests and native Bayes and classifies songs to predefined emotions. Another is Song Level Dimensional MER uses regression such as support vector regression, and gaussian process regression to classify the song into different dimensions (such as valence and arousal). By using these algorithms, the machine is able to classify the emotion based on the musical piece.

3.2 HIERARCHICAL FRAMEWORK CLASSIFICATION

The paper "Automatic Mood Detection and Tracking of Music Audio Signals" Lu et al. (2006), is an example of how this system works. Musical features such as tempo, loudness, pitch change were automatically extracted through MIDI Files, converted into frames containing spectral bands. Analysis using the Fourier transform Frequency domain converted the images to features such as timbre (based on spectral shape) and intensity (based on subband energy). The information was then fed into a hierarchical framework, which then classified the music by emotion (contentment, depression, exuberance, and anxiousness). By using a hierarchical framework, the music features known to be more impactful in setting the mood were able to be weighed more heavily in musical analysis. As a result, 800 short clips of acoustic, classical music with a testing accuracy of 86.3 percent.

3.3 Deep Learning Neural Networks

The paper "Music emotion recognition using recurrent neural networks and pretrained models" Grekow (2021) utilizes deep learning in training machines to recognize the sentiment that music brings. In contrast with the traditional machine learning method, no specific musical aspects need to be extracted prior to running it through the algorithm. Instead the system will automatically extract suitable features depending on the data. This is important to bring good results, as often the different quality and methods of extraction of different musical pieces can lead to differing results. Through deep learning, convolutional or recurrent neural networks (CNN and RNN) can be used as an end to end processing framework. CNN Models are frequently used, as it can learn feature representations based on data effectively.

3.4 Bi-Modal Deep Boltzmann Machine

Another example of utilizing Deep Learning in MER is outlined in the paper "Bi-Modal Deep Boltzmann Machine Based Musical Emotion Classification" Huang et al. (2016). The Bi-modal Deep Boltzmann Machine architecture (DBM) is a deep neural network which bases its probability on the Boltzmann Distribution and is based on the Restricted Boltzmann Machine. It uses two layers of DBM networks (one for audio, one for lyrics) as well as one additional layer to join the two. Once the machine has selected suitable features of the given track, these features are inputted into SVM, as used in traditional machine learning, and classified into the suitable categories. The result of this proved to outperform other singular modality models, which proves that the DBM model is effective in determining music sentiment due to its consideration of the relationship between lyrics and audio features.

3.5 Visual Geometry Group

Visual Geometry Group Net, an improved version of CNN was used in paper "Recognition of emotion in music based on deep convolutional neural network"(Sarkar et al., 2020) to explore the performance of different audio features to the emotion it evokes. Due to the increased number of layers, music recognition accuracy was increased relative to CNN models however the model struggled to identify arousal and time series nature of audio could not be properly represented in the model (See Sarkar et al. (2020) for more information)

4 Data Processing

4.1 Data Sourcing

To prevent potential violations of copyright infringement, our group made an effort to collect samples from royalty-free sources. By using the DEAM dataset, compiled by researchers from the University of Geneva in 2015, we were able to do so. These samples came from the archives of "freemusic.org", "jamendo.com" and the Medley DB dataset. The set contained 2000 samples of royalty-free music in MP3 format, collected in a 45-second excerpt, clipped randomly throughout a song. The truth labels of the dataset were provided useing Russel's Circumplex Model, as mentioned in the background works, evaluating each song on a degree of valence and arousal, ranging from 1 to 9 (See **?**. for more information). Based on these scores, the mood of the song can be found.

4.2 Data Pre-processing

To clean our data, we had to ensure there was a variety of moods in our data set. First, the moods of "Happy", "Sad", "Calm" and "Aggressive" were defined based on the statistical mean of our dataset, with a value of 4.5. The threshold bounds can be seen in Table 1 for our four different emotions, along with the data set distribution based on class. To ensure a more equal representation of mood classes, we selected 223 samples from each class, with an exception of 203 from the "Sad" class due to lack of samples. This is important because if there are significantly more samples in one category, our model will be highly trained in identifying songs in the specific category while doing poorly in other classes. By ensuring a more even distribution of samples, we have avoided creating a biased model.

Table 1: Sample Distribution based on Class Valence and Arousal Values

Mood Class	Arousal Score	Valence Score	Number of Samples
Happy	Greater than 4.5	Greater than 4.5	223
Sad	Less than 4.5	Less than 4.5	223
Calm	Greater than 4.5	Less than 4.5	203
Aggressive	Less than 4.5	Greater than 4.5	223

The data was processed into three unique features, each which is researched to be impactful on the sentiment it brings. These three features provide a reasonable level of complexity while still capturing fundamental dimensions of music closely correlated with genre as well as the emotions it is likely to evoke. The features are the following:

1. **MFCCs (Mel-Frequency Cepstral Coefficients):** These coefficients capture the short-term power spectrum of sound, providing information on the variety of timbre texture in the song, and distinguishing between different tonal qualities. To extract the MFCCs, we used the librosa.feature.mfcc function which calculates the MFCC over a timeframe based on the inputted audio array, sample time, MFCC coefficients, number of samples in each short-time fourier transform frame and it's hop length. We set the sample time to 44100 seconds, with 13 MFCC coefficients, 2048 short-time fourier transform frames and a hop length of 512.
2. **Spectral Contrast:** This extracts the difference between peaks and valleys in a sound spectrum, aiding in the differentiation between timbral tones; it offers clues about the harmonic and non-harmonic content of a song. Spectral contrast was extracted using the spectral contrast function in Librosa's features based on the inputted audio array, short-time fourier transform frame of 2048, hop length of 512, minimum frequency of 200 Hz, and seven bands. Bands were set to seven as this is the highest and most detailed for this sound spectrum.
3. **Chroma:** This feature provides a snapshot of the energy distribution across twelve pitch classes, corresponding to the twelve musical notes in an octave. It represents harmony, chord progressions and tonal structures.Chroma can also be extracted using Librosa's feature Chroma function. The function outputs chroma represented in tensor form based on the inputted audio array, sample time of 44100, short-time fourier transform frame of 2048, hop length of 512 and chroma of 12, corresponding to the 12 pitches in Western music (See Han et al. (2022)).

Since the outputted dimensions of the features are not all the same in length, padding needs to be added. This results in a 13 x 3800 x 3 matrix, capturing both frame-level representation and sequential data. One sample of the training or validation data would contain two tensors, with the first tensor as the music feature data, and the second tensor which represents the truth label as a size 2 array with it's valence and arousal numbers. A cleaned data sample's features can be visualized in the figures below.

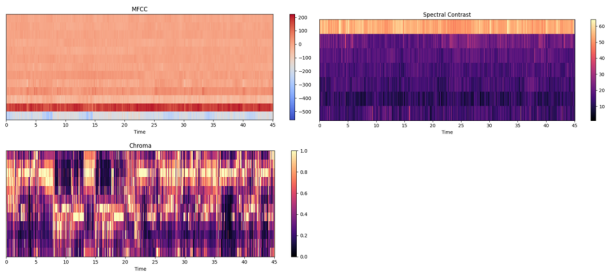

Figure 2: Cleaned Sample Data Visualized

5 ARCHITECTURE

Our model architecture is a CNN with two convolutional layers and pooling layers followed by three fully connected layers, all with ReLU activation function. CNNs are great for pattern detection from spectrograms which is precisely the input format we will gather in our data-processing stage. The idea is to use the CNN layers for extracting spatial features and temporal patterns from audio features. Once the important features and patterns are identified and learned, the fully connected layers will help combine the learned features together, leading to the correct classification. Our model also uses drop out for regularization to prevent over fitting of the model at the second last fully connected layer.

In terms of the heterogeneity of the parameters, we have adopted a feature stacking method, where the data is normalized and concatenated so that no feature initially overpowers the other. This will ensure the input data is a hierarchical classification scheme that could be implemented to improve classification accuracy by breaking down the classification task into simpler, more manageable tasks, which could lead to more accurate and interpretable classification outcomes.

6 BASELINE MODEL

As our baseline model, we selected the Support Vector Regression (SVR) which is a supervised learning algorithm, commonly used for classification tasks. SVR predicts continuous numerical values by finding a hyperplane that best represents the relationship between the input features and the target outputs. For this task of music emotion classification, we will use the earlier mentioned features to find the best fit of a non-linear hyperplane which results in a low mean squared error (MSE) value and high R-squared score (R^2) value.

The simplicity of the baseline model allowed us to set a performance benchmark for the more complex model and ensure that the primary model can outperform the SVR and yield outputs with less error. It also helped resolve fundamental issues with loading the data and proved that the problem itself was solvable using machine learning. For simplicity, we only used the first three inputs from the extracted MFCCs to obtain the valence and arousal scores.

Based on the results from the progress report, the best scenario for valence was (RBF, with C value = 10) while the best scenario for arousal was (POLYNOMIAL, with C value = 10, and Degree =3).

A qualitative note on the performance of the model was that the error varied significantly based on the size of the dataset. A smaller dataset resulted in a higher error while a larger dataset resulted in a smaller error.

A challenge encountered while using the model was that the SVR does not take in empty datasets or datasets with a value of NaN because it cannot be scaled based on the set margins. To overcome this, the data was filtered first before being input into the SVR model.

7 QUANTITATIVE RESULTS

Quantitative assessment of our model's performance over the training and validation data was made by utilizing loss, and mean absolute error (MAE) metrics. Briefly, loss quantifies the penalty a model suffers from a bad prediction, and MAE is a measure of the difference between the model's prediction and the corresponding ground truth, fitting for our model whose predictions can be mapped to a two dimensional space, and the difference is very easily visualized (similar to a euclidean distance). We expand on this visualization to understand model performance in section 8: *Qualitative Results*

Figure 3 and 4 are graphs of loss and MAE over the course of training (number of epochs) of our model. We can clearly see the continuity and more specifically a significant downwards trend of both loss and MAE across both training and validation data sets.

The exact values for each metric our model, which were chosen among the lowest metrics of various competing models were recorded (see Table 2). Note these are the metrics we obtained after the last epoch of training, or in other words the final metrics. As previously stated, our models metric values

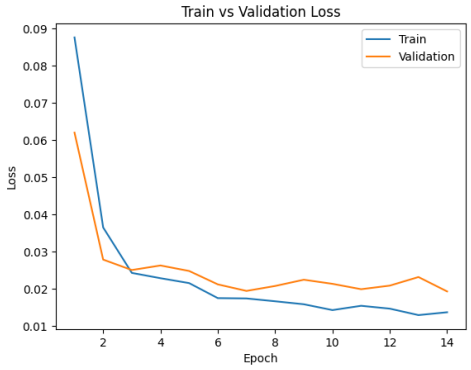

Figure 3: Model training/validation loss.

Figure 4: Model training/validation MAE

Table 2: Values of training and validation metrics

Metric	Dataset	Value
Loss	Training	0.0136
Loss	Validation	0.0192
MAE	Training	0.0905
MAE	Validation	0.1108

are among the lowest of which we tested, and are clearly lower than our baseline model's equivalent metrics. This demonstrates the superior efficacy of our model as opposed to comparable models.

We can make many assertions using this data that may not explicitly present themselves at first. For example, our model has a very low loss rate compared to MAE. This means that while it may not be incredibly precise in determining the exact arousal and valence scores, it does not suffer for this because it is able to correctly classify the overarching emotion of the model. This means that if our model were to be used practically, it would be successful due to the fact that nearly all tasks it would be used on would be more concerned with the classification of a song's emotion, and not how precise it computes internal scores used to output said emotion. Considering we are successful on the classification side establishes its usefulness in practice, and demonstrates that high MAE can be misleading without context.

8 Qualitative Results

Obtaining a genuine and palpable understanding of our models performance by looking at graphs of various metrics is very difficult. To make this easier, we developed, and present the outputs of a function that visualizes our models performance used on our model in figure 5. To be specific, the function chooses a few predictions our model made at random to plot, as well as their corresponding ground truth label as two distinct points. We express this difference the two points by plotting a line between the two. This enables us to obtain a deeper qualitative understanding of our models performance.

Looking at the function's outputs in figure 5, we can establish with certitude our model's efficacy in identifying valence and arousal scores for different musical scores, and thus the songs emotion. This is because of the easily observable small distances between model prediction and ground truth, indicated by the length of the line that connects them. Shown in this figure although, is an example of our model performing slightly worse on certain groups of data. As we can see, the lengths of the lines are longer in quadrants excluding the bottom left, which corresponds to sad songs, whereas lines connecting predictions and ground truths both in the sad quadrant are extremely close. We diagnose this as our model overfitting to songs classified as being in this bottom left quadrant, thus tuning its weights with a bias towards outputting predictions in this region. This hypothesis is

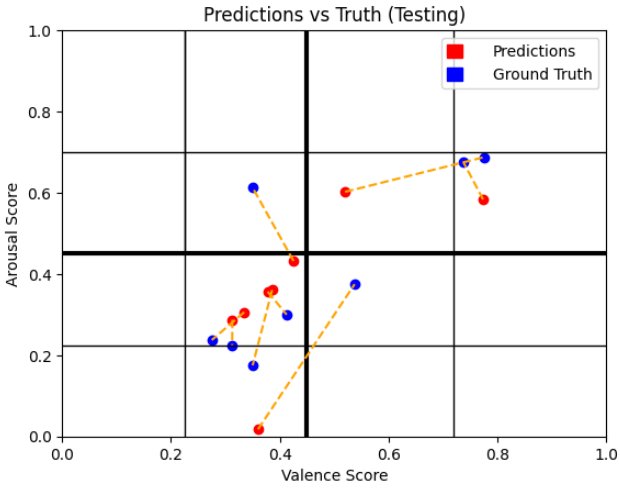

Figure 5: Plot of model prediction and corresponding ground truth label

further supported by the lengths of the lines mentioned, as there are several examples of the model predicting a song is sad, when its ground truth is in other quadrants, and in addition, the model is extremely precise in classifying sad songs.

During data processing, we made sure that each of the 4 main quadrants/emotions had an equal sample size of 223. This would suggest that a model that has weights balanced to each emotion would have a similarly equal spread of predictions when shown the entire dataset, but this is not what is exactly shown in figure 6.

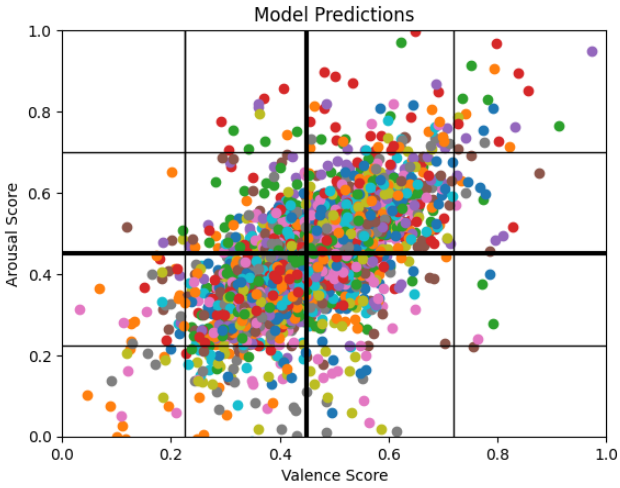

Figure 6: Distribution of models predictions over training data.

We can clearly see the model exhibits a slight but tangible bias towards the sad quadrant. This means that weights potentially have a slight bias to this quadrant which supports our assertion above. This results in some samples being classified as sad when they may actually be in the other 4 quadrants, meaning a slight imbalance to sad predictions.

Overfitting and unbalanced weights to certain classes will likely pose an issue going forward with the model that is worth solving. To that end, we propose reducing our batch size. What this does is allow our model to ”look more closely” at each sample, allowing it to capture robust features and develop stronger, more precise weights. We also propose using early stopping, that would allow

us to use more balanced weights computed earlier on in a training session rather than weights that slowly become more biased to a certain class, deeper into the training.

9 Evaluation of Model on New Data

To ensure our model works correctly and that the test results show a good representation of the model's performance on new data, the team had to obtain brand new samples for which the model had not been trained on. Our training data samples were collected in 2015 from sites such as "freemusic.org", "jamendo.com" and the Medley DB dataset. So, to ensure the samples were brand new and had not been seen by the model, the team made sure to collect royalty-free 45-second track snippets from the site "chosic.com". Additionally, to further ensure there were no repeated samples from the training song set to the test set, the team selected test tracks which were created after 2015. As a result, our test data is a brand new dataset, never on training our model before.

Additionally, the new samples were given in 45-second MP3 files and pre-processed to have it's features presented in the same format as the training, identical to how the model would be used in real-world situations. However, the new test data samples did not have any valance and arousal scores, which is needed in order to quantify the error between the predictions of the model and its correct answer. As a result, valence and arousal scores of the new music data were assigned from manual rankings done by the team. Consideration of the music tags posted on "Chosic.com" was also used when assigning truth values. Each team member listed to a sample and ranked the sample from 0-1 on valence and arousal. Then, the average of the team ranking was used and taken as its truth value. The team acknowledges this system of manual ranking of the test dataset comes with some biases and potential error, as it is only our four team members who have ranked our test data. However, due to our time and resource limitations, it was the most feasible truth values as we do not have access to a large team of people who can help us rank music.

To further test the abilities of our model, the team considered testing the model with pieces of music which had been slowed down, sped up or put into reverse to potentially trick and better train the model, similar to how a picture is stretched out in an image identifying model. However after further discussion, it was concluded that since the relationship between music and emotion is extremely time sensitive, altering the model relative to time would completely change the sample, creating a new music piece with a different truth label. Thus, testing the model with brand new music samples was sufficient method for testing.

Additionally, the team ensured there was an equal distribution of samples for each class. This allows the model to be tested on its performance on each class metric. For test data, 20-23 samples from each class was collected, totalling 93 new data samples.

10 Discussion

Overall, the model does a good job of assigning the correct labels to the inputted data features based on the performance metrics used. From the new dataset of our hand-collected and self-evaluated test, our model performed well by accurately modelling the test music to its true mood class. Once the test set was ran through our machine, it had a loss of 0.063577 and MSE of 0.214. The test results have higher error than our training and validation results which is to be expected.

One reason for our increase in error may be due to the fact that our truth values for our test set was manually evaluated by only 4 people, where as in typical study, these values would be reviewed and evaluated by everyone.

The limitations of our model are multifaceted and influenced by several key factors. Firstly, the inherent subjectivity of music significantly impacts the model's ability to accurately define emotions. This subjectivity introduces errors in how our model interprets and defines emotions within musical pieces. Additionally, the size of the dataset and the granularity of defined emotions play a crucial role in the model's performance. A larger dataset with more refined emotional sub classes tends to yield better results. Furthermore, certain emotions may be more prevalent in music than others, which can bias the model towards recognizing these emotions more accurately compared to less common ones. Finally, the complexity of musical content is another challenge, as music can often convey multiple emotions simultaneously or transition between various emotional states. Accounting for

this multifaceted emotional content adds an additional layer of complexity to the model's accurate interpretation.

While the model already considers sequential data, there is an opportunity to delve deeper into sequential patterns by leveraging Recurrent Neural Networks (RNNs) and extending the sampling time per sample. Additionally, the significance of a song's lyricism and musical symbols cannot be understated. Enhancing the model's performance is possible through sentiment analysis applied to these elements. With a larger data set, through transfer learning we can also create a better-performing model, however with transfer learning, it is difficult to tell which samples have been used for training for testing purposes, giving a less accurate test.One could also argue the prevalence of certain musical emotions in music. Stronger emotions warrant a stronger need to produce music or for music listeners to seek out.

11 Ethical Considerations

Many music tracks today are copyrighted, and proper usage rights and licensing must be obtained in order to be used for machine learning legally. To avoid violating usage rights and copyright law, the team has used data from royalty-free and public-domain music for our testing, training and validation data. This ensures proper licensing for our research purposes.

However, the team acknowledges potential bias from only using royalty-free and public-domain music. Different kinds of music exist in all cultures. Since we collected sample sets from websites with English writing, our samples are reflective of English-speaking cultures. Furthermore, music is a subjective art form. Different cultures have different interpretations of music. By having a limited data sample, our network will was trained on data sets from English-speaking cultures. By having this limitation, our data could potentially misidentify emotions music brings because it lacks cultural context.

12 Project Difficulty and Quality

The project we have undertaken presented a very high level of complexity and difficulty. This can be identified through the advanced nature of our data processing methodology, our model architecture, as well as the subjective nature of the task at hand. In utilising sophisticated spectrogram features instead of relying solely on the audio frequencies from the raw files, we adeptly translated complex audio data into a format that was richer in detail and information, more conducive to a successful deep learning model. The use of Convolutional Neural Networks (CNNs) in our model architecture was a particularly key decision; we harnessed the CNN's ability to capture spatial characteristics of audio data and extended it to a traversal over the time axis, as per the design of our kernels in the convolutional layers. This was instrumental in accurately recognising emotions, and this idea came as the product of our knowledge of the dynamic and temporal nature of emotional interpretation in music, which we have previously discussed.

Another significant challenge in this project was the inherent subjectivity associated with emotional responses to music. This subjectivity presented substantial obstacles in data labelling and necessitated a nuanced approach to training, validation, and testing methodologies. To this end, we employed a credible and rigorous, quantifiable metric: namely the use of arousal and valence values. From these values we developed a multi-levelled approach to the interpretations of results, as due to the inherent subjectivity, the interpretation itself a challenge. This approach provided two levels, where we analysed the ability of models to first identify basic emotions, and in terms of more abstract and subjective emotions. Additionally, we reorganised our dataset to ensure a balanced representation of emotional states due to a concern regarding model bias towards emotions that were more prevalent.

Comparatively, the project's model performance was superior to the baseline Support Vector Regression (SVR) model, as evidenced by lower loss and mean absolute error (MAE) metrics. This achievement is particularly noteworthy considering the challenging nature of this project. Therefore, our team not only tackled technical difficulties that are inherent in deep learning programming, but also demonstrated a deep understanding of and capability to manage the subtleties of subjective data interpretation.

REFERENCES

J. Grekow. Music emotion recognition using recurrent neural networks and pretrained models. *Journal of Intelligent Information Systems*, 57(3):531–546, 2021.

Donghong Han, Yanru Kong, Jiayi Han, and Guoren Wang. A survey of music emotion recognition. *Frontiers of Computer Science*, 16(6):16, 2022.

Moyuan Huang, Wenge Rong, Tom Arjannikov, Nan Jiang, and Zhang Xiong. Bi-modal deep boltzmann machine based musical emotion classification. In *Artificial Neural Networks and Machine Learning – ICANN 2016*, volume 9887, pp. 199–207. Springer, 2016.

Lie Lu, Dan Liu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):5–19, 2006.

R. Sarkar, S. Choudhury, S. Dutta, A. Roy, and S. K. Saha. Recognition of emotion in music based on deep convolutional neural networks. *Multimedia Tools and Applications*, 79(1-2):765–783, 2020.