⟨Dissertation Title⟩

⟨Student Name⟩

MSc ⟨EngD⟩ in ⟨Programme⟩
The University of Bath
⟨Academic Year⟩

⟨Dissertation title⟩

Submitted by: ⟨Student Name⟩

Copyright

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Masters of Science in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Abstract

⟨ The abstract should appear here. An abstract is a short paragraph describing the aims of the project, what was achieved, and what contributions it has made. The purpose of the research (what's it about and why's that important) The methodology (how you carried out the research)) The key research findings (what answers you found) The implications of these findings (what these answers mean)⟩

Contents

List of Figures

List of Tables

Acknowledgements

Add any acknowledgements here.

Chapter 1

Introduction

1.1 Music Emotion Recognition (MER)

1 page TO DO Music Emotion Recognition (MER) is a subset of the broader field of Music Information Retrieval (MIR); a discipline which is dedicated to the extraction and analysis of information from music data (Ma et al., 2024).

1.2 Background and Context

2.5 page

TO DO

1.2.1 Defining Emotion and its Role in Music

TO DO

An Overview of Music Information Retrieval (MIR) and its Significance

TO DO

Objectives and Scope of Music Emotion Recognition

TO DO

The Significance and Applications of MER

TO DO

1.3 Problem Definition and Current Challenges

2 page

1.3.1 Emotion Representation Models in MER

TO DO

Categorical Models (Discrete Emotions)

TO DO

Dimensional Models (Valence-Arousal Space)

TO DO

Hybrid Emotion Approaches

TO DO

1.3.2 Principal Challenges in MER

TO DO

1.3.3 The Need for Multimodal Approaches

TO DO

1.4 Research Questions and Objectives

1 page

1.4.1 Primary Research Question

TO DO

1.4.2 Secondary Research Questions

TO DO

1.4.3 Project Objectives

TO DO

1.5 Approach and Expected Contributions

1 page

1.5.1 Methodological Approach

TO DO

1.5.2 Expected Deliverables

TO DO

1.5.3 Potential Applications

TO DO

1.6 Dissertation Structure and Organisation

0.5 page

"Music can lift us out of depression or move us to tears – it is a remedy, a tonic, orange juice for the ear. But for many of my neurological patients, music is even more – it can provide access, even when no medication can, to movement, to speech, to life. For them, music is not a luxury, but a necessity." Oliver Sacks, best-selling author and professor of neurology at NYU School Of Medicine. cite here

make sure all these points are covered - statement of the aims and objectives of the project what you are going to deliver as evidence that you have reached those objectives. There should also be an indication of the context in which this problem exists (its background and history) why this is an important problem to consider.

state what you are going to be discussing in your dissertation, and why the problem that you have proposed is worth to be academically investigated to the point that you can discuss it in a quite competent way, using technical terminology.

You should include a discussion of related work (work by other people), where it has bearing on your own work. This gives a context for your project and will help you in your critical evaluation of your achievements. The related work will be referenced in your text, and the references detailed in your reference list..

You should include a survey of the background and historical overview of your project topic, to demonstrate why it is an important problem to consider and to show your engagement in researching and understanding the literature.

In some cases, you will want to discuss why others' attempts to solve the same (or similar) problem have failed, or why you are choosing to follow one person's approach above all the rest.

Also indicate how you intend to extend or incorporate what you have learned from others' work into your project. Remember to reference all the literature that you survey in the reference list.

This section should include an initial description of the problem that you are investigating, its background and history. This will give a context for the main work of the investigation.

It is quite likely that your project will be rather specialised, so you will need to ensure that you provide sufficient information for the reader of your dissertation to be able to understand and appreciate the topic. It is very important to thoroughly research the background of the problem you are investigating, so that you can be clear about the aims and extent of your investigation. Ensure that the reader knows exactly what you intend achieve.

NOTES application in - e.g. recommendation and therapy. customisation of material such as songs and recommendations, by understandint eh emotional context of a oerson a system can better align with this and provide a suitable selection of music. This can also be used in film scores and automatic music choice in advertisements (increaing advertising efficacy) , filems, and any othe presenation context. monitoring in mentak health and also interesngly used in memory aids for people with alheimers etc. can be used in identifying mood disorders and treatment (FIND REF HERE THIS SOUNDS INTERESTING AND WHAT I WANT TO HELP CREATE). Also helping to categorise music as maintaining metadata is impossible as growing content with the increaed peoduction of music due to AI assistance and tech helping, plsu people using tools to produce music that wouldn't have previously been possibke. think orchestra back in the 1800s fr example. Retreival and categorisation can be helped by MIR in general otherwise hidden music stays hidden. Ineresting future developments are emotion recognition agents that can then b used to play a particukar sone or example a song on a phone ap using pyshiological cues or facial cues. might not aways want to stay in that state. maybe want to change the state such as starting where you are then improving with music.

Good quote - These days, music permeates practically every situation we find ourselves in, including those involving our daily activities like eating, sleeping, cleaning, shopping, studying, exercising, and driving [30]. That said this is an older paper - has anyone said anything similar R.E. Thayer, "The Biopsychology of Mood and Arousal", Oxford University Press, 1990.

The explosion of digital music has dramatically changed our music consumption behavior. Massive music libraries are available through streaming platforms, and it is impossible to browse the entire collections item-by-item. As a result, we need robust knowledge management systems more than ever. oaprphase from won textbook

reasons it is inportant 0 can use content based recommendation, can use AI to curate opkaylists online, can use music to categorise new music rather than i being dne manually ANOTHER GOOD QUOTE Since almost all musical compositions aim to evoke a particular feeling in the listener, music has been categorized and retrieved according to emotion [32][53] [32] Z. Xiao, D. Wu, X. Zhang, and Z. Tao, "Music Mood Tracking Based On HCS", In 2012 IEEE 11th International Conference on Signal Processing (Vol. 2, Pp. 1171-1175), IEEE, 2012. 53] V.A. Kumar, C.V. Rao, and N. Leema, "Audio Source Separation by Estimating the Mixing Matrix in Underdetermined Condition Using Successive Projection and Volume Minimization", International Journal of Information Technology, 15(4), 1831-1844, 2023.

ANOTHER GOOD QUOTE Studies on music information behavior suggest that individuals also consider emotions when choosing and organizing music [37] 37] C. Huang, D. Shen, "Research on Music Emotion Intelligent Recognition and Classification Algorithm in Music Performance System", Scientific Programming, 2021, 1-9, 2021.

mention this here [50] J.A. Russell, "A Circumplex Model of Affect", Journal of Personality and Social Psychology, 39(6), 1161, 1980.

DIMENSIONAL MODEL - emotions are defined as quantitative values along different dimensions of emotion

valence (referring to the pleasantness or unpleasantness of emotional states), potency (representing dominance or the perception of control and freedom to act), and arousal (indicating the level of energy and stimulation) [13][23]. AGAIN REFS ARE NOT RELEVANT 13] Y.H. Yang, Y.C. Lin, Y. F. Su, and H.H. Chen, "A Regression Approach to Music Emotion Recognition", IEEE Transactions On Audio, Speech, and Language Processing, 16(2), 448-457, 2008. [23] D. Wang and X. Guo, "Research on Intelligent Recognition and Classification Algorithm of Music Emotion in Complex System of Music Performance", Complexity, 2021, 1-10, 2021. what is this potency that he mentions???

I intuit that there are more than 2 dimensions and anything that we use otherwise is not useful as we are losing a dimension or two and flattening. I have found a paper for 8D emotion scale. It would be interesting to read this and test out whatever I create on this mapping. Or at least a 3d model that includes a dimension for emotional not emotional.

Also sound is movement. it is liternally movement of air.

categorical approach too

psychologists frequently rely on people's verbal reports of their emotional responses. How does it make you feel? But it might make you feel something but you know the feeling is something else. eg the different types of emotion need to re-look back into this.

Universal across cultures, basic emotions are commonly linked to specific physiological changes or emotional expressions [14]. the ref doesn't seem to be right here but 14] K.W. Cheuk, Y.J. Luo, B.T. Balamurali, G. Roig, and D. Herremans, "Regression-Based Music Emotion Prediction Using Triplet Neural Networks", In 2020 International Joint Conference on Neural Networks (Ijcnn) (Pp. 1-7), IEEE, 2020. Also used adjestive clustering done by people can research this

However again I feel this isnt right. has someone done experiments into embedding on emotion? LOOK INTO THIS.

Quote Eeriola here. do i have this A. Gabrielsson, E. Lindström, "The Influence of Musical Structure on Emotional Expression", 2001. make a comment here or have a rabbithole into emotion on the side. i dont think emotion is 2d. t all. i think it has many dimentions. maybe look into papers on this as the 2d modelseems as aflat as pants music.

I am watching a video onhow music affects the brain. high frequenciies affect hairs at the base of the cochlie whereas lower frequencies at the apex, higher amplitude means more firing, overtones affect a compleax patterm, clochlear nuclei process sound first and interpret the sound goes therough a few other unconscious bits then hits auditory cortex, which has an area for raw pattern and other belt areas that link to other areas. processes auditory scene anakysis. travels to other areas. and signals flow back and forth from these areas. areas are - intervals, meldoy and harmonay are processed, rhythm from movement area like cerebellum, meaning syntax and structure overlap with language processing area in the brain. emotions affected by limbic system, auditory network links to this too, limbic helps with memory too so hence research in alzheimers, music activates reward system, drop helps with dopamine!! I did not know this. synchronise helps people bond too.

music as medicine

Hevner emotion loop, - Lok into this - was mentioned in Lin 2025. I haven't heard of it, ALSO MENTIONED IN kang are we there yet review paper. Also geneva music scale which is categorical but has more tags 45 tags and 9 categories. mtg Jamenda dataset which has free tags from listeners so technically categorical,

most common approach RTussels circumplex model of effect. Discuss Valenca (positive negative) arousal (active or passive) and how the originated from the circumplex model. There is another dimension dominance but not used. This is harder to agree on - quoted on Kang paper 2024 are we there yet.

Thayer model too to discuss. energetic and tense arousl - not used as much.

for principle challenges cover these points The Subjectivity of Emotional Perception

data Scarcity and the Annotation Bottleneck

The Semantic Gap Between Low-Level Features and High-Level Emotions

Chapter 2

Literature and Technology Survey

2.1 Methodological Approaches to Unimodal Music Emotion Recognition

3.5 PAGES. Unimodal approaches to MER have relied on either audio data or lyrical text analysis Yang et al. (2023).

2.1.1 Early Audio-Based Methods

Audio-based MER attempts to predict human emotion directly from the acoustic signal itself, and rely on signal processing techniques to extract the relevant features from the audio. Classical Machine Learning (ML) techniques focussed primarily on feature engineering, which required expert domain knowledge to handcraft emotionally salient features Louro et al. (2024). (Yazhong Feng, Yueting Zhuang and Yunhe Pan, 2003) carried out the first Audio-Based MER experiment, using two musical features, tempo and articulation, to predict a classification of emotion into 4 classes, achieving a precision and recall of 67% and 66% respectively.

TO DO

Feature Extraction and Representation

TO DO

Conventional Machine Learning Paradigms

TO DO

Deep Learning Approaches

TO DO

previously emotions are labelled and a classifier learns to categorise them.

REFER TO GLOBALMOOD (LEE) PAPER HERE AND MENTION THE DIVERSITY OF DATASETS AND CULTURAL DIFFERENCES IN LABELLING MUSIC Difficulty is inpointing which parts of the music make the emotional context of a song. Is it the combination? Is there a recipe? Where you need the basics but anything else is superfulous. I dont think so. support vector machines, have been used to examine which characteristics are importnat. [12] R. Panda, B. Rocha and R.P. Paiva, "Dimensional Music Emotion Recognition: Combining Standard and Melodic Audio Features", In 10th International Symposium on Computer Music Multidisciplinary Research–CMMR 2013 (Pp. 583-593), 2013. read this and see if relevant - i defo also have some papers on MIR and feature extraction.

dataset s are also subjective (copy the parts from my proposal here)and hve bias etc.

intro to the dufferent nethods that have been used to do both the classification task and the feature extraction.

mention that some things are categorised differently depending on culter Undoubtedly, there exists a multitude of relationships that remain undiscovered, irrespective of whether their dynamics are influenced by cultural norms or not [48].

Also good to mntion here the different types of emotion incud and percieved whithc can further complicate the area. can quote gomez here.

Subjectivity of emotion - highly personal and can change - copy the part from my propsal here

Ambiguity of the eemotion labels - the choice between discrete and continuous emotion scale intordoces ambiguiuty. categorical might be too small a grouping, wheres dimensional may not show the complexity of msuci. mention the 7 dimensional proposal by that music paper on emaotion model. There has been some effort to map these - find reference here.

Feature representation is a huge part of this - what features are relevant, what methods are prime for getting these

model generalisability is a big issue. it trained on one data set most models di not generalise well. this is due to different musical styles, recording confisions and annotiation methodologies. This is also a problem in SSL read this and reference it GENERALIZATION OF SELF-SUPERVISED LEARNING-BASED REPRESENTATIONS FOR CROSS-DOMAIN SPEECH EMOTION RECOGNITION Abinay Reddy Naini

Interpretability - when using deep learning

Data scarcity and quakity (e.g., MFCCs, chroma, rhythm features, spectral features). Deep learning-based feature learning (e.g., raw audio, spectrograms). Traditional audio featu mention mel spectorgrams being similar to human ear and recognises weights the parts of audio that are detectable by humans. mention here is anyone uses dufferent here in any papers and cover ant discussion points on whether this is best. can also pick this up in the discussion. look at preemphases on a to see if this is needed apparently older funciont Pre-emphasis: for the feature extraction make sure you go into detail for the hop length, n mels and the nfft because in the methods section I have written that I have written about this here. This is the only step that *librosa.feature.melspectrogram*() does not perform by default. Pre-emphasis is a more traditional technique, and with modern deep learning models, it's often considered an optional or unnecessary step. The process of converting the spectrogram to a decibel scale

$$(usinglibrosa.power_t o_d b)$$

, which you are already doing, helps to balance the spectrum in a similar way.

DICUSS USING THE NP GLOBAL MAX AND WHY THIS DIDN'T WORK, CAN USE SONG MAX BUT THEN EACH SONG IS NORMALISED TO ITSELF AND YOU LOSE DATASET

DIFFERENCES IN AUDIO. I OPTED FOR 1 WHICH IS THE DIGITAL STADARD LOOK AT AN INTRODUCTION TO AUDIO CONTENT ANALYSISAdvantages of ref=1.0 - This made my plots look odd so went for global max.

MENTION GRADIENT CLIPPING AS AN OPTION FOR WHEN USING 2 DIFFERENT MODELS IN CNN NOT REALLY NEEDED BUT WITH BERT AND CN IT IS RESEARCH THIS THIS I WHAT I IMPLEMENT.ED

ALSO FOR ATTENTION MODULE THERE ARE LOTS OF DIFFERENT TYPES OF ATTENTION - SU 24 PAPER IS GOOD READ THIS AND GET NOTES. I PIKED A BASIC ONE WAS GOING TO DO SQUEEZE AND EXCITE BUT NOT SUR IF REALLY NEEDED THE PAPER THAT QUOTED IT WASN'TGOOD.

2. Frame segmentation and windowing: Yes, this is done automatically. The function breaks your audio into short segments (frames) behind the scenes. You can control this with parameters like n_fft (the frame size) and hop_length (the amount the window shifts for each frame). It also applies a window function (by default, a Hann window) to each frame to prevent spectral leakage, just as described.

3. Calculation of the magnitude spectrum (DFT):

Yes, this is done automatically. This is the core of the Short-Time Fourier Transform (STFT) which melspectrogram runs internally to get the frequency information for each frame.

4. Calculation of the Mel-spectrum (Mel-filterbank):

Yes, this is the main purpose of the function. After calculating the standard spectrum, it applies the $Mel-filterbank$ to convert the linear frequency scale to the perceptual Mel scale. You can control the number of Mel bands with the n_mels parameter.

mention here that the tokeniser also tonises the ounctiation. whether this is sensible for a lyrics model im not sure. you dont sing punctuation in the same way that you say it.

Also history of it - Consequently, some intriguing findings emerged, such as the fact that major modes are typically linked to emotional states like joy or solemnity, while minor modes are connected to negative emotions like melancholy or fury [1][2][20][38]. before machine learning, [20] R. Malheiro, R. Panda, P. Gomes, and R.P. Paiva, "Emotionally- Relevant Features for Classification and Regression of Music Lyrics", IEEE Transactions on Affective Computing, 9(2), 240-254, 2016. this is an interesting paper I have ND WANT TO READ, [30] R.E. Thayer, "The Biopsychology of Mood and Arousal", Oxford University Press, 1990.

cover feature based methods too

I particularly find it hard to break a song down into features. Its the combination. Could you recreate a song using the same feature values that has the same score on an emotional scale when evaluated either by a model or a person?

Has anyone done raw audio? If so, use this paper.

straightforward, consonant harmonies are typically upbeat, relaxing, or pleasant. Contrarily, because they produce instability in a musical work, complex, discordant harmonies are associated with feelings like anxiety, excitement, or melancholy [19][28]. feature extraction relevant here.

timing, dynamics, articulation, timbre, pitch, interval, melody, harmony, tonality, and rhythm

For feature extraction tecnicques has anyone looked at what our brain does? Does music affect animals? Mosquitos listening to dubstep eat less and have less sex!!! A 2015 study conducted at the University of Wisconsin-Madison found that while cats are happy to ignore regular 'human' music, they are highly responsive to music that is written especially for felines. Cows produce more milk when listening to slower music. Kennelled dogs are happiest when listening to soft rock and reggae

pre set rules eg major is happy. kang also mentioned that hierarchical framwork to map acoustic features to physchological models e,g., thayer,

* 2.1. Traditional Approaches * 2.1.1. Acoustic Feature Engineering (e.g., MFCC, chroma, spectral centroid, tempo, rhythm, harmony, timbre, dynamics

explain why deep learning became necessary or superseded some of these classical methods, providing a smooth transition. You touch on this in your notes, but make sure it's explicit in the text. Feature engineering such as Timbre, Ryhtm, Harmonics and Dynamics of the song. These features were handcrafted, Ince the feaures were identified calssical mchine learning models were used. could bt SVM, GMM, and KNN, and random forest. liitations were this was reliant on domain expertise, and the methods didn;t model the elements well and performance plataued. Quote jiantg here Music Emotion Recognition Based on Deep Learning: A Review and * 2.1.2. Classical Machine Learning Models (e.g., SVMs, GMMs, Random Forests).

CNNS Convolutional Neural Networks (CNNs) for Spectrogram Analysis.

RNNS Recurrent Neural Networks (RNNs, LSTMs, GRUs) for Temporal Modeling.

TRANSFORMERS Transformer Architectures in MER.

HYBRID . Hybrid Models (e.g., CRNNs

cnn - patterns lima louro A Comparison Study of Deep Learning Methodologies for Music Emotion Recognition, using g mel spectropgrams, which are a 2d image that represents the features of the music over time the NNs they learn the hierarchical features of the music using these 2d representations that relate to timbre, harmonay and rhythm. .

rnn - lstmt for changes over time. WORK Intemporal dependacies over time so can see how parts of the song unfold over time. Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure can also quote hAN HERE. MUSIC EMOTION RECOGNITION AND CLASSIFICATION USING HYBRID CNN-LSTM DEEP NEURAL NETWORK CAN QUOTE DUTTA BUT NEED TO READ IT FIRST THIS PAPER IS RECENT THOUGH FROM 2024

transoformers applied too now. these started off being used for NLP but have gained populatiry in this area and other audio processing. see hiw parts of the song ate to each other for example building tenisoion in a peice may lead to a different emotion later one. self attention mechanism enables them to odel long range dependencies and contextual relationships in the musical sequences more effectively. review by liang covers this

A lot of models have used a combination so hybrid models. so for example using cnns for feature extraction and rnn for temporal modelling.

mention that kang says not really a clear winner in sota as there is so much cariation in the datasets used and emotion modelling. most reecntly most researchers are using their own dataset.,

2.1.2 Lyric-Based Methods

TO DO

Text Representation Techniques

TO DO noes rom AI - The Hugging Face Tokenizers library is the industry and research standard for this task. By identifying it, you are looking at the correct, state-of-the-art approach. et's look at the steps you found. They are all essential parts of the modern NLP pipeline:

Breaking down text into tokens: The Hugging Face tokenizer does this using advanced subword tokenization (like WordPiece for BERT). This is more powerful than just splitting by spaces because it can handle rare words and variations of words (e.g., "sing," "singing," "sang") more effectively.

Assigning a unique numerical identifier ($input_ids$): This is correct. After tokenizing, every token is mapped to a unique number from the model's vocabulary. This is the first step in turning text into numbers the model can process.

Converting to a tensor: This is also correct. The list of $input_ids$ is converted into a tensor (like a PyTorch or TensorFlow tensor), which is the required input format for deep learning models.

Adding special tokens: This is a crucial step specific to models like BERT.

[CLS]: This token is added to the beginning of every sequence. The model uses the output corresponding to this token as a summary of the entire sentence, which is perfect for classification tasks like predicting emotion.

[SEP]: This token marks the end of a sequence or separates two different sentences.

Traditional NLP Methods (e.g., Bag-of-Words, TF-IDF)

Word Embeddings (e.g., Word2Vec, GloVe)

Contextual Embeddings (e.g., BERT, GPT-like Models)

Classification Models for Lyrical Analysis

TO DO Models for text classification (e.g., Naive Bayes, SVMs).

Deep learning architectures for lyrics (e.g., RNNs, LSTMs, Transformers).

Challenges Specific to Lyrical Content

TO DO (e.g., poetic language, slang, context).

2.1.3 Evaluation Metrics and Benchmarks for Unimodal MER

TO DO Kang - metrics used to evaluate are dependant on what dataset is used. categorical models - Accuracy - how often is it right, precision - and area under the curve. measure of how weel the model can distinguish between different abels.

dimensional is a regression task. use - mean squared error, r squarede, and pearson correlation.

Hard to compare models as there are different datasets with different annotation methods, labels, emotion type and modelling, song length, genre, size etc so it becaomes a really difficult task to compare f different models use different datasets. comparing apples and oragnges. mirex, medieval. mentioned in kang as they will try to standardise it with competitions etc. apparently the aera has declined to less benchmarking possible.

2.2 Multimodal Music Emotion Recognition

2.5 pages Multimodal MER (MMER) expands on uni-modal approaches by leveraging information from sources beyond audio. These modalities include physiological data such as Electroencephalogram (EEG), textual or symbolic notations like lyrics and music manuscripts, visual data from music videos, and digital symbolic representations like the Musical Instrument Digital Interface (MIDI). By combining these complimentary information sources, the aim is to construct more nuanced models that achieve higher accuracy in recognising emotional content in music.

2.2.1 The Rationale for Multimodality in MER

TO DO Multimodal models have been shown to consistently ourperform unimodal models (Liyanarachchi, Joshi and Meijering, 2025)

. Why combining modalities is beneficial (complementary information, robustness).

Limitations of unimodal approaches.

2.2.2 Multimodal Fusion Strategies

TO DO

Early Fusion

TO DO . Concatenation of features before input to a single model.

Advantages and disadvantages.

Late Fusion

TO DO . Separate models for each modality, then fusion of their predictions.

Advantages and disadvantages.

Hybrid/Intermediate Fusion

TO DO . Fusion at various layers within deep learning architectures.

Discussion of common architectures (e.g., your proposed CNN-Transformer fusion).

Examples from existing literature.

2.2.3 Datasets for Multimodal MER

TO DO . Review of prominent datasets (e.g., MediaEval, EMOTIC, DEAM, and specifically MERGE, if it has existing literature).

Discussion of their characteristics, annotation schemes, and limitations.

The importance of large-scale, well-annotated datasets. MAYB MENTION SOME OTHERS AND THEIR ISSUES BEFORE DOING A WHOLE SECTION OR PARAGRAPH ON MERGE. MERGE IS SUPPOSED TO BE DESIGNED WITH THESE ISSUES IN MIND SO USE THIS PAPER FOR THIS SECTION

2.2.4 State-of-the-Art in Multimodal MER

TO DO . Detailed review of recent significant works in multimodal MER.

Highlighting models that use similar architectures to yours (CNN for audio, Transformer for text).

Analysis of their reported performance metrics and methodologies.

this may not be the case if i dont do similarity

2.3 Cross-Modal Analysis and Similarity Learning in MER

(2 pages -

2.3.1 The Concept of Cross-Modal Congruence

TO DO

2.3.2 Similarity Learning in Deep Learning

TO DO

2.3.3 Applications of Similarity Learning in Multimodal Contexts

TO DO

2.3.4 Prior Research on Cross-Modal Relationships in MER

TO DO

2.4 Music Emotion Datasets and Data Preparation

2 pages -

2.4.1 A Survey of Prominent MER Datasets

TO DO

2.4.2 Supervised Datasets for Model Training and Evaluation

TO DO

2.4.3 Corpora for Self-Supervised and Weakly-Supervised Learning

TO DO

2.4.4 The Importance and Challenges of Cultural Diversity in Datasets

TO DO

2.4.5 Data Preprocessing and Normalization Techniques

TO DO

2.4.6 Data Augmentation Strategies

TO DO

2.4.7 Inherent Challenges in Music Data Curation and Scale

TO DO

Mention here - this is mentioned in the merge dataset paper, tha valencce and arousal are not predicted equally. audio struggles with valence. audio is good at predicting arousal. but not valence. find references here but lyrics are better at predicting valence outperform. words provide context. this is why merge was created. merge is yhe biggest bimodal dataset. they averaged the values and transposed them from the all music ratings. is this ok - what were the allmusic tatings from? they were checked after, but still is it easier to agree when you know what the previous outcome was? was it done blind? if not it should have been. need to look more into this mapping dictionary and whether it is ok - how it was constructed. the dataset also cuts out the middle parts - e.g. the not strong ones. This seems concerning to me, as hiw can you build a picture with obscuring some parts f tyhe data? i get they were trying to train on sronger candidates, but for me that seems not great.n its like learning about neon colours then the model would struggle with slight hints. not sure about this dataset.

tHERE ARE TWO MAIN TYPES OF DATASET (kAND 2024 ARE WE THERE YET) - Static and dynamic. Staic gives one label for en entire song whereas dynamic notes the changes in a piece as it goes along and reocds how this changes.,

MTG - Jamenda - stats (kang)

Moodswings - kang. has dynamic labels.

MERP have both static and dynamic which is helpful.

Also when looking at datasets needs to consider whether they are balelling based on induced or perceived emotion, induced is how it makes you feel and perceived is how you think the song is categorised. The latter is more common as there is less objectivitiy. i.e. a happier song could make you feel sad if it reminds you of somenone who has passed, but you can still say that is objectively a happy song.

most datasets look at perceived emotion whereas sme look at indcued. - eeg, skin conductance, ecg, etc. DEAP, HKU956, mUSIC-mOUV, SiTunes.

Induced and perceived are connected but not identical.

Most are audio files but the paper (Kang) mention MIDI. Emopia is a midi dataset, pnda et als dataset and VGMIDI, ym2413-mdb and popular hooks.

Some datasets are muti modal some have lyrics and video.MuVi has music videos DIQ as well. DMDD, popular hooks and merge have the lysics too. RAVDESS has speech and music with emotional exressions. Might be usefull to see if this is good as a pretraining task before training on music. Some thngs in music are based on sounds we hear from other people and tones they use.

Kang review says that adding lyrics doesn't actually help increase score that well which is surprising.

Kang says datasets are an issue as they are small compaed to other fields, Also some have ehole songs and others have clips so not standardised. mrp, vgmidi USE FULL SOnggs, some are 30 seconds to 1 mns like DEAm and emopia. mer 500 uses 10 second peices.

Different genres vary between the different datasets. DEAM - rock and electronic DEAP - pop, rock, classical, and jazz vg-midi - video game sound tracks. some dont state - e.g. mtg jamendo, music4all, muse. mer500 hindi film music. CCMED-WCMED - western and chinese classical music

Labeled MER Datasets for Fine-tuning/Evaluation (Overview of key datasets like PMEmo, DEAM, MTG-Jamendo, EmoMusic, Soundtracks; focus on their characteristics: size, audio format, emotion labels, annotation process, availability)

Large-Scale Unlabeled/Weakly-Labeled Datasets for SSL Pre-training (e.g., AudioSet, FMA, MagnaTagATune, MSD, NSynth; their suitability for learning general music representations)

The Importance and Challenges of Culturally Diverse Datasets (e.g., limitations of Western-centric datasets, efforts like GlobalMood).

Audio preprocessing techniques Discusses data considerations and preprocessing requirements for ssl

Data augmentation strategies - Examines technical challenges and potential solutions specific to SSl

Challenges specific to music data - sixe of datasets etc?

Really good paper to quote is Watcharasupat 2025, quote (Music Emotion Recognition Music emotion recognition (MER), also known as mood recognition, is an archetypal example of a task without an absolute ground truth.) which mentions that some things dont have ground truth and dont have good rate of interrater agreements. These should instead be treated as a distributions. This means that we need to consider uncertainty quantification. Uncertainties in

Machine Learning (ML) systems are commonly categorized into data (aleatoric) uncertainties and model (epistemic) uncertainties — quote this need to get the paper Mucsányi, B., Kirchhof, M., Oh, S.J.: Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. In: 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada (2024)

Nead to read up on this paper here for the Watcharasupat - He says that deep learning systems have been designed as determenistic maps and are not well suited for tasks where the ground truth is not the same as noise. He also states that the older papers look at the ground truth as a distribution however more recent and neural network based mer studies look at the mean and median of the labels to get the answer when this isn;t an appropriate thing to do. He makes a good point that we are starting to use MER in therapy so we need to be sure we are using the right data.

Annotation process as mentioned in Kang review - free tag approach used in mtg-jamendo sometimes a lit. Sometimes crowdsourced, smaller groups e.g. music experts.

label distribution in datasets means that the datasets are not balanced - may have lots of some tags but hardly any odf things that are less common eg, scARED. This means its harder for the AI to pick up on this and train. DEAP has tried to be more balanced.DEAM says that the test train split is defined.

2.5 Pervasive Challenges and Open Research Problems in MER

1.5 pages)

2.5.1 Subjectivity, Label Noise, and Ambiguity in Musical Emotions

TO DO

2.5.2 Optimizing Pretext Task Design for Emotion-Relevant Features

TO DO

2.5.3 Domain Adaptation and Generalization Challenges

TO DO

2.5.4 Model Interpretability

TO DO

2.5.5 Computational Requirements

TO DO

2.5.6 Real-time Processing Considerations

TO DO

2.5.7 Evaluation of SSL Representations Beyond Downstream Accuracy

TO DO

. Subjectivity, Label Noise, and Ambiguity in Musical Emotions (Impact on SSL fine-tuning and evaluation

Optimizing Pretext Task Design for Emotion-Relevant Features (Ensuring SSL tasks learn musically and emotionally salient information)

Domain adaptation issues -Cross-Corpus, Cross-Lingual, and Cross-Cultural Generalization (Domain adaptation, performance on diverse music

Model interpretability - Interpretability of Learned Features in SSL-MER Systems (Understanding what SSL models learn about music and emotion).

Computational requirements

Real-time processing considerations

Evaluation of SSL Representations Beyond Downstream Accuracy (Probing tasks, clustering quality, visualization techniques for MER)

2.6 Future Directions and Outlook

Future Directions and Outlook (Approx. 1 page

2.6.1 Emerging Trends in MER

TO DO

2.6.2 Advancements in Music-Aware Pretext Tasks and Architectures

TO DO

2.6.3 Integration with Other MIR Tasks

TO DO

2.6.4 Ethical Considerations and Computational Efficiency

TO DO

Emerging Trends in SSL for MER (e.g., Multimodal SSL, personalized/context-aware MER, few-shot/zero-shot learning, continual learning, integration with generative models

Advancements in Music-Aware Pretext Tasks and Architectures

Integration with other MIR tasks

Potential applications

Ethical Considerations and Computational Efficiency.

2.7 Conclusion

Conclusion (1 page)

2.7.1 Synthesis of Key Findings from the Literature

2.7.2 A Critical Analysis of Current Approaches

2.7.3 Future Directions and Open Challenges in the Field

Summary of key findings

Critical analysis of current approaches

Concludes with future directions and open challenges in the field design marking criteria

MARKING CRITERIA This is the chapter in which you review your design decisions at various levels and critique the design process. if the project involves investigatory or experimental work then you will need to report how you chose and designed the experiments and discuss the criteria by which your experiments will be judged to be successful or otherwise (benchmarking). You should describe, and justify your choice of experiments, explaining clearly what sort of information they are intended to provide and how that information will be used. Then you can detail the design of the experiments. When you design your experiments, it is essential that you state the criteria by which your experiments will be judged successful. How will you evaluate the performance of your experimental software? Justify all your choices in design/implementation, and try to relate this part to insights that you may have acquired from the literature and technological review. The detailed content of this part of your dissertation will depend very much on the specific project you are undertaking, and it is up to you to decide the best way to report your work. You should discuss this with your project supervisor. This should include details of scientific or technical problems that had to be tackled, including clear descriptions of these problems how you approached solving them, your proposed solutions to these problems, and a justification of why your proposed solutions are appropriate

Section 3 - Self-Supervised Learning (SSL) as a Paradigm for MER (Approx. 2.5 - 3 pages

heading - Fundamental Principles and Benefits of SSL (not realted to MER)

Advantages over supervised approaches

NOTES

cover here how this has worked in image and how it can be applied generallt to udio including MIR tasks. mention how a spectrogram is made which relates to the image part and the pros and cons of this being the informatio stoarge for music. mention that some have worked on audio instead.

how ssl is Addressing Data Scarcity in MER

Learning Robust and Generalizable Representation

NOTES

This is from AI so not finalised but will go along something like this. I'll know more once I know if it is possible to implementPrimary Research Question "How can X-Sample Contrastive Learning be effectively adapted for Music Emotion Recognition to improve performance over traditional contrastive learning approaches?"

Secondary Research Questions "To what extent does incorporating emotion similarity relationships in the contrastive learning objective improve emotion recognition accuracy compared to binary positive/negative designations?" "How do different formulations of emotion similarity metrics (dimensional vs. categorical vs. hybrid) affect the quality of learned representations in music emotion recognition?" "Does X-CLR demonstrate improved data efficiency for MER tasks compared to standard contrastive approaches, and if so, under what conditions?" "How effectively can X-CLR disentangle emotional content from other musical attributes (genre, instrumentation, production style) in the learned representation space?" "What temporal modeling approaches best complement the X-CLR framework when capturing emotion dynamics in music over time?" "How does the temperature parameter in the X-CLR similarity softmax function affect the balance between emotion precision and generalization capabilities?" "Can X-CLR learn meaningful cross-cultural patterns in music emotion recognition, or does it primarily reinforce culture-specific associations?" These questions provide a comprehensive framework that covers the technical adaptation of X-CLR to audio data, the effectiveness of various emotion similarity definitions, performance improvements, and specific capabilities that would be valuable for MER applications.

I'd recommend selecting 3-4 of these questions as your core focus, ensuring you have a manageable scope that still allows for meaningful contributions. You could also refine these based on the specific papers you've been reviewing and any particular aspects of music emotion that most interest you. mention There is a scarcity of superior, standardized datasets and due to the issue with labelling being biased, thenself supervised makes sense. also music can chagne within a song. mention some trained on video may not understand irony and the purposeful jaxtaposition of music and a scene to create a moment,

Also need to understand a bit about the people and their preferences. Do people like metal rate jazz. FOR MERGE DATset i Noticed all the metal was in one quadrant. Is this right? Do people vote a certain way for music that doesn't appeal to t them?

SSL aims to ccapture meaningful representations of the music by leveraging the inherant structure and information without relaying on any labelling of the data. The meaningful patterns in the music can be leveraged to complete pretext tasks and the mdel learns by solving these yasks. In doing this it captures features that can be useful fordownstream tasks.

As these work on unlabelled data - it helps solve the problem of good quality labelled ata. It aLO AIMS to kearn the mid level features such as timbre etc that might be used to preduct emotion, however also learns higher level musical features and sematics which can be useful for MER.

issues these days according to Kang are as follows 1. Limited dataset limityations - we dont have ebough large diverse, datassets that dont suffer from copyright issues. lots are genre sepcific so dont generalise well enough. self supervised may be the answer here.

2. Subjectivity of labels and cultural biases 3. noisy labels 4. annotation dynamic interfaces - hard to collcet the data. mechanicla turk may provide incorrect data so need to check interrater reliability 5. no clear benchmarking and difficult comparing. a mapping has been devied (I was going to do this!) which was y !@ Seeing stars of valence and arousal in blog posts by Georgios Paltoglou and Michael Thelwall . in affective computing. referenced in Kang.

Weights oif neurak ebtworks need to be initiailised befire traiing commences. need to try and train at a global minimum. can use a large labelled dataset to help choose weights. this is called pretraining not great though if the thing is biased or likely to be incorrect as in the emotion peice. train on unlabelled aata and test on abelled. still an issue if labels are wrong,using same weights, mention here To explore how self-supervised learning can address the challenges of limited labeled data in MERa and To evaluate the potential for creating more robust and generalizable MER systems

Overview of SSL Techniques for Audio/Music

Contrastive Learning (e.g., SimCLR, CPC, Myna for music; pretext tasks, augmentations). NOTES

Contrastive Learning - learns representations where the different augmented views of the ausio smaple are pulledcloser togethw in the embedding space - examples include SIMclr. CLMR, Myna, SimCLR adaptations the audio version on this is CLMR and Myna. However this can impact the information and emotion so this is critical what augmentation take place. quote andread Myna: Masking-Based Contrastive Learning of Musical Representations Ori Yonay Contrastive Learning from Synthetic Audio DoppelgängersManuel Cherep

also i wasn't ging to use ssl but COULD I MYABE? AND REPEAT THE CNN PART WITH A SSL METHOD AND COMPARE THESE?

Predictive Learning (e.g., APC, future frame/masked part prediction

NOTES

Predictive coding - This task is where the the model predicts future segments of an audio sample.Such as contrastive predictive coding (CPC) and Autoregressive Predictive Coding (APC). Compares its prediction to the actual future part of the music. it is given the true and a part that isn't true. quote Audio-Based Emotion Recognition Using Self-Supervised Learning on an Engineered Feature Space nimitsurachat

Masked Modeling (e.g., Masked Autoencoders like AudioMAE, BERT-style for audio).

other SSL methods for audio - e.g. the below - maybe lose this section if not enough to fill it with,

NOTES

and there are some on hugging face too. motion-Anchored Contrastive Learning Framework for Emotion Recognition in Conversation Published on Mar 29, 2024Authors:Fangxu Yu although more about audio not just music/

Temporal jumbling/Swapping of audio. n example is the Audio-only Self-Supervision (Odd) method, where 25 percent of clips are jumbled, and within these, two windows are swapped; the encoder then identifies the swapped elements referenced in minanchurat paper

Predicting Engineered Features - he PASE (Problem Agnostic Speech Encoder) predicts things like waveform,

can also use generative models where you recreate the origianl audio from one which has added noise 0 aautoencoders, denoising autoencoders.

nOTES

Masked Audio/Spectrogram Modeling inspired by VERT and masked languaged modelling, mask out portions of the spectrogram and the model has to reconstruct this. This is the idea behind A-JEPA and AudioMAE. A-JEPA, AudioMAE, MERT, MATPAC. same namachiruat referenceas above for this. also read and quote S3T: SELF-SUPERVISED PRE-TRAINING WITH SWIN TRANSFORMERFOR MUSIC CLASSIFICATIONHang Zhao1 and motion2vec: Self-Supervised Pre-Training for Speech Emotion RepresentationZiyang Ma1

Generative Modeling (e.g., Autoencoders, VAEs, VQ-VAEs for audio; role of latent spaces).

NOTES

Schubert investigated a few of the relationships in Russell's emotion model between these characteristics and the emotional reactions [48]. [48] Y. Wang, S. Sun, "Emotion Recognition for Internet Music by Multiple Classifiers", In 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS) (Pp. 262-265), IEEE, 2019. ref not correct here again.

Section 4 - Prominent SSL Models and Architectures in the Context of MER (Approx. 2 - 2.5 pages)

NOTES

Need to mention here that a lot of papers (need to find out how many as I am going through this, assume independance of the 2d relationship bbetween valenance and arousal. This is a like,ly incorrect assumption. another reason why self supervision might be better. can mention some studies here where ereseaxchers have used full covariance matrices in their work. (so not assuming independance)

NOTES

Id there a such thing as a model which uses a generator model to product music and this is evaluated by the MER extraction method utilising self supervised methods. The generator could be API from a well known model. Could tweak some of these features if this is possible.

Transformer-based SSL Models for Audio

Speech-derived models (e.g., Wav2Vec2, HuBERT) and their applicability/transferability to MER

Music-specific Transformer models (e.g., MERT) and their use in emotion-related tasks

Vision Transformers (ViTs) for Audio Spectrograms (e.g., AST, Myna – architecture, parameters, efficiency

Joint-Embedding Predictive Architectures (JEPA) for Audio

The I-JEPA Paradigm (Context/Target Encoders, Predictor, Latent Space Prediction, EMA)

A-JEPA for Audio Spectrograms (Adaptations, Masking Strategies like Curriculum Masking, Regularized Masking

Stem-JEPA and other JEPA-style audio models (e.g., from Sony CSL, MATPAC

Other Relevant SSL Frameworks (e.g., SSSL, emotion2vec, EmoGen – briefly, if space allows, as examples of diverse SSL applications in emotion/music) Cross-modal learning

NOTES

Transformers are increasingly prevalent especailly for masked modelling and ocontrastive learning, Their self attanetion is good at long range dependencies, in sequential data. Models like Wav2Vec2, HuBERT, and MERT are prominent SSL Transformers for speech and music. Vision Transformers (ViTs) are used as backbones in A-JEPA and Myna when processing spectrograms as images. Towards Unified Music Emotion Recognition across Dimensional and Categorical Models Jaeyong Kang and EAT: Self-Supervised Pre-Training with Efficient Audio Transformer Wenxi Chen and Exploring Acoustic Similarity in Emotional Speech and Music via Self-Supervised Representations Yujia Sun,

CNNs are still important but mainly used for feature extraction from spectograms. used alongside transformers mainly. MERTECH: INSTRUMENT PLAYING TECHNIQUE DETECTION

USING SELF-SUPERVISED PRETRAINED MODEL WITH MULTI-TASK FINETUNING Dichucheng Li 1

RNNs ar not used as much but can be used sometimeswith modelling temporal aspects or in hybrid models - Semi-Supervised Self-Learning Enhanced Music Emotion Recognition Yifu Sun

Teacher student architectures - Teacher-Student Architectures like A-JEPA. There is a smaller student model which learns to mimic a larger teacher model the target enocedoer which is the teacher has its weights as an exponential moving average of the context encoder (student) weightsm which provides stable learning targets. MATPAC model also uses this student teacher relationship. SSLAM: ENHANCING SELF-SUPERVISED MODELS WITH AUDIO MIXTURES FOR POLYPHONIC SOUND SCAPES Tony Alex, MERT ACOUSTIC MUSIC UNDERSTANDING MODEL and this Literature Review] Masked Latent Prediction and Classification for Self-Supervised Audio Representation Learning Masked Latent Prediction and Classification forSelf-Supervised Audio Representation Learning Aurian Quelennec

WITH LARGE-SCALE SELF-SUPERVISED TRAINING Yizhi Li and A-JEPA: Joint-Embedding Predictive Architecture Can Listen Zhengcong Fei,

muQ - Zhu - uses MELRVQ. pre processing step (tokenizer) which takes mel spectrogram and makes sequence of simpler units or tokens. uses a random projection quatiser. as other random starting point can affect the output in the end. bestRQ is what ws used before. encodec is another thing that was used but resource intensive. MelRVQ is already pretrained on music data beforehand, uses 3 losses. codebook loss. typical musical sounds. helps buil core musical dictionary. comittment loss. trains encoder to make sure output maps to the dictionery. reconstruction loss. trains decoder to be able to recreate me spectrogram to recreate the same mel spectrogram from the reconstruction based on simplifications. losses are balanced at the end. understnds musical structure. melrvq creates the same sequnce/tokens over again with diffrent level of detail. Also different prediction heads used in MuQ different perspctives. simple architecture so less labour intensive than ncodec. melrvq is the teacher and muq is the student. conformer network. worked well - outperformed mert and musicfm. acheived this with a lot less training data so more efficient. broad range of MIR not just emotion. blation studies showed removing pretrainnig and this was vital. music mulan is a model that uses lyrics (text) and how thi relates to the music. did a layer analysis and it shows acoustic features were learned in lower layers and modd etc wa in later layers, so shows a hierarchcail learning is taking place.

Chapter 3

Design

In this chapter, I will outline the iterative development of the research methodology and provide a design justification for the methods I employed. Adaptations were made as a direct response to challenges uncovered while training on the MERGE dataset, a novel multimodal MER dataset. With the exception of the introductory paper by the dataset's creator, Louro et al. (2025) there is no established research precedent for this dataset and its applications in MER research.

TO DO - ADD ANY OTHER CHANGES TO THE DATASETS OR METHODOLOGY HERE IN INTRO IF NEEDED

3.1 Computational Workflow and Infrastructure

The following sections outline the computational setup and choice of methodologies for the data loading and preparation, which were optimised to overcome performance issues.

3.1.1 Experimental Setup: Hardware

The experiment was carried out on two platforms: a local MacBook Pro M1 for CPU-based tasks and a cloud-based Nvidia A100 GPU for model training. To compare performance between the local CPU and the cloud GPU, a preliminary benchmark test was conducted using the standard Colab Nvidia T4. The time taken for a 10-epoch training loop using simulated data and a simple CNN was recorded and is detailed in Table 3.1.

Table 3.1: Performance Benchmarks Across Platforms

Task	Local Mac (CPU) Time	Colab (GPU) Time
Audio Processing (Librosa)	∼22.18s	∼15.94s
Audio Processing (Torchaudio)	∼0.05s	∼0.76s
CNN Training (10 epochs)	2.92s	1.20s

3.1.2 Data Handling and I/O Optimisation

To overcome I/O bottlenecks that were encountered during both data preprocessing and training, at the start of each session the dataset was copied into the local filesystem of Colab's runtime environment. This eliminated the network latency that was experienced when reading the files from the mounted Google Drive, as data could be accessed directly from the virtual machine's high-speed local storage.

3.2 Dataset and Preprocessing

3.2.1 The MERGE Dataset: An Overview

The MERGE dataset was chosen for this dissertation which, at the time of writing is the largest publicly available bimodal MER dataset. This was published recently by Louro et al. to address the "severe lack of public and sizeable bimodal databases", and as yet, there have been no published papers using MERGE. Furthermore, the curation process was rigorous, incorporating a manual validation step, ensuring data quality. MERGE is provided in both a "Complete" and "Balanced" version. The "Complete" version was chosen for this dissertation, as it contains more data points, rendering it more suitable for a regression task. To ensure methodological consistency, MERGE also offers two predetermined Train-Validation-Test (TVT) splits: a 70-15-15 split and a 40-30-30 split. The 70-15-15 split was chosen to leverage the larger training set, providing the model with more examples to improve generalisation.

A significant challenge in MER research is the processing of large audio files, which can create computational and I/O bottlenecks during model training. The MERGE dataset mitigates this issue by providing audio in 30-second MP3 files, reducing data handling requirements. The metadata, such as artist, year and genre, is also made available where possible, so further analysis can be undertaken if required.

3.2.2 Audio Signal Processing

Before being fed into the CNN, each audio file was converted into a Log-Mel Spectrogram, a 2D representation of the audio, using the following data pipeline:

1. **Loading:** Audio files were loaded using the Librosa library at a sample rate of 22,050 Hz, a standard rate for MER tasks Chaturvedi et al. (2022). The MERGE paper's authors (Louro et al., 2025) used 16,000 Hz to reduce the complexity of the model, stating that "such reduction does not impact the model's performance [6], as confirmed experimentally". However, a closer examination of their reference revealed that Pyrovolakis, Tzouveli and Stamou (2022) only compared downsampled values from 44,100 Hz to 22,050 Hz and not 16,000 Hz. For this reason, I opted for the more widely-used 22,050 Hz sample rate.

2. **Mel Spectrogram Generation:** A Mel spectrogram was computed using a Short-Time Fourier Transform (STFT) with the following parameters: an FFT window size of 2048, a hop length of 512 samples, and 128 Mel bands. These parameters are a common choice for MER tasks Pandeya and Lee (2024) and the detailed reasoning is provided in the previous chapter.

3. **Decibel Scaling:** The resulting Mel spectrogram was converted to a logarithmic decibel scale, which helps to balance the dynamic range of the audio and mimic human auditory perception Ma et al. (2024).

4. **Standardised Length:** To ensure a uniform input size for the CNN, all spectrograms

were normalised to a fixed length of 1292 time-frames. Spectrograms longer than this were truncated, while shorter ones were padded with -80 dB (approaching silence) at the end. As the majority of the samples were 30 seconds in length, this standardisation only corrected minor length variations within a small portion of the samples, and ensured uniform input dimensions with minimal information loss. One outlier song was found that was 4.5 seconds long, which was likely an error in the dataset. However, the decision was made to keep this sample in the experiment, as, on listening, there appeared to be some emotional content present.

5. **Normalisation:** The decision was taken to not normalise the audio, as volume is a relevant feature that requires preserving. However, a form of normalisation takes place when converting the audio to a spectrogram, as the conversion requires a maximum power value as a reference point. As discussed by (Cazzaniga, Gasparini and Saibene, 2024), a global maximum power was first calculated and all spectrograms were scaled to this maximum, preserving the unique characteristics of individual songs.

The final output of this preprocessing stage was a 2D tensor of shape (128, 1292) for each audio track, serving as the input to the convolutional feature extractor.

3.3 Core Architecture: Bimodal Valence-Arousal Regression Model

The foundational model for the first two experiments was a hybrid deep learning architecture designed to predict Valence and Arousal values. The model was composed of two towers: a Convolutional Neural Network (CNN) for the audio, and a Transformer-based model for the lyrics. The outputs of both of the streams were fused before being passed to the final prediction head. CNNs have been used extensively in MER tasks as they mimic the visual perception of humans by learning hierarchical spatial patterns and can learn feature representations effectively Su et al. (2024). Recurrent Neural Networks (RNNs) are also commonly used in MER, however their strength lies in their ability to capture sequences over time and are more often used in dynamic emotion detection Han et al. (2022). Since the chosen dataset only provides one single static emotion label, a CNN was chosen.

3.3.1 Audio Feature Extractor: Convolutional Neural Network

The initial plan was to replicate the CNN architecture described in (Louro et al., 2025) however, upon implementation it was found that there were several inconsistencies in the dimension handling and feature merging. The figures appeared to contain notation errors, such as mixing 1D and 2D operations for sequential text data and inconsistent pooling specifications, which suggested potential transcription errors in the publication. As it was unclear which architecture was used, I decided to opt for another method instead.

A VGG-style Convolutional Neural Network (CNN) was therefore chosen as this approach is widely used in the field Won, Spijkervet and Choi (2021) and has been shown to be effective for various audio tasks Wang et al. (2024). The design is inspired by the VGGNet architecture, where square filters capture visual associations across the two orthogonal dimensions, and can be applied to the 2D image of a spectrogram Fong, Kumar and Sudhir (2025).

Audio Feature Extractor: Model Architecture

To evaluate different architectures, an ablation study was conducted on five different models, with different architectures. The performance of all five models is shown in Table 4.1. A VGG-inspired CNN (Model 4) enhanced with an Attention Module, Batch Normalisation and Dropout Regularisation was selected as the feature extractor. Additionally, training optimisation techniques were implemented, including early stopping to prevent over-fitting and learning rate scheduling to improve convergence. The network consists of four sequential convolutional blocks. Each block is designed to capture increasingly complex patterns from the spectrogram input:

- **Block 1: Low Level Feature Extraction** The initial layer was a block of two consecutive 2D convolutional layers, each with 64 output channels, each with 3x3 filters, designed to learn the low-level features like edges and gradients from the spectrogram. Each convolutional layer was followed by Batch Normalisation, which was added to accelerate training convergence and improve stability. It can also help with regularisation leading to suppression of over-fitting Han, Chen and Ban (2023),Qiao et al. (2024). An Activation Layer followed each Batch Normalisation layer, to introduce non-linearity to the model. A ReLU Activation function was chosen, as this is one of the most widely used activation functions, aiding faster convergence and assisting in combatting the vanishing gradient problem Yang, Liu and Gong (2025). Finally, the outputs were passed to a 2x2 Max-Pooling layer. This layer downsamples the feature map by taking the maximum value from each 2x2 window and discarding others, which decreases computational load and provides translational invariance so that the patterns themselves are recognised regardless of their local position. Zhao and Zhang (2024).
- **Block 2: Mid-level Pattern Recognition** This block adopted the same overarching structure with two consecutive 2D convolutional layers, each with 128 3x3 filters. Each layer was both followed by Batch Normalisation and ReLU and the entire block finished with Max-Pooling.
- **Block 3: High-level Structure Detection** Similarly, this layer contained two consecutive 2D convolutional layers,with 256 3x3 filters, each followed by Batch Normalisation and ReLU with the block terminating with Max-Pooling.
- **Block 4: Feature Abstraction** The final feature extraction layer contained two consecutive 2D convolutional layers, each with 512 3x3 filters and followed by Batch Normalisation and a ReLU activation. The block finished with an Adaptive Average Pooling layer. This layer takes the spatial average of each feature map, and ensures the output is a 512-dimension vector, preparing it for the classifier.
- **Classifier Block: Dimensionality Reduction and Regularisation** The classifier transforms the 512-dimensional feature vector through a series of layers, each with progressive dimensionality reduction. The first linear layer reduces the features to 256 dimensions, followed by ReLU activation. Dropout layers with a 50% probability are applied before the first linear transformation and before the Attention Module to prevent over-fitting. An Attention Module is applied to the 256-dimensional features to weight their relative importance. Finally, the second linear layer produces the final 64-dimensional feature vector.

Ablation Study

- **Model 1: Baseline Lightweight VGG-Style Architecture** The initial baseline model was a simplified VGG-inspired CNN with 4 convolutional blocks. Each block consisted of a single convolutional layer followed by a ReLU activation and Max-Pooling. The channel dimensions were the same as described in the model above.
- **Model 2: Batch Norm & Regularisation** This model enhanced Model 1, by the addition of Batch Normalisation. A Learning Rate Scheduler and Early Stopping were implemented for athis model and all subsequent versions.
- **Model 3: Feature-wise Attention** The third iteration of the CNN contained an attention module that followed the final 64-dimensional feature vector from the fourth block. The attention module consisted of two layers which mapped the 64 dimensions down to 16 then back to 64, with both ReLU and Sigmoid activations applied.
- **Model 5: Unfrozen BERT** The fifth model utilised the same pre-trained BERT model however its weights were 'unfrozen', allowing fine-tuning during training. In previous iterations of the models, weights had been previously frozen to increase speed and reduce computation.

3.3.2 Lyrical Feature Extractor: Pre-trained Transformer Model

TO DO For the text data a pre-trained transformer-based approach was used to extract the semantic features. BERT (Bidirectional Encoder Representations from Transformers) base-uncased model was chosen

3.3.3 Bimodal Fusion and Regression Head

TO DO

3.4 Experiment 1 Baseline Performance on the MERGE Dataset

TO DO The first experiment aims to establish a robust performance baseline using the recently published MERGE dataset The protocol is designed for reproducibility.

3.4.1 Dataset Specification and Preprocessing

TO DO

3.4.2 Training Protocol and Evaluation Metrics

TO DO

The final output of the audio tower is this dense, 64-dimension feature vector, which is engineered to encapsulate the essential emotional characteristics of the audio track, ready for fusion with the features extracted from the lyrics.

decided on a basic vgg ish from advice in the won textbook 2021. fo small models. i looked and 4 papers (see the emotions echo for reference, that 4 papers use a vggish model

i decided to use a pretrained model instead and used a transformer for the yrics part although i did use their cnn archtecture for the audio tower.

i need to use the freezig approachch du to training times as this was taking too long fnie tuning the whole thing. quote

For context, a different study by Malheiro et al. (2018) using lyrics-only regression achieved an R 2 score of 0.59 for arousal and 0.61 for valence, suggesting that lyrics analysis alone showed much better performance for valence than similar audio-only studies at the time (which obtained 0.28 for valence) ADD THE ABOVE IN. BECAUE IF MOST OF THE ANALYSIS USES TEXT AND AUDIO BRINGS IT DOWN ARE WE NOT THEN BETTER OFF JUST DOING TEXT WHERE THE SIMILARITY IS HIGH AND ADJUSTIG THE WEIGHT OF EACH BASED ON THIS????

3.5 Experiment 2: Comparative Analysis on the [Second Dataset Name]

TO DO The second experiment investigates the generalisability of the model and explores the impact of dataset characteristics on prediction behaviour. The same core architecture from Section X.1 is applied to a different, widely-used dataset in MER research.

3.5.1 Dataset Specification and Preprocessing

TO DO

3.5.2 Justification for Comparative Analysis

TO DO

3.6 Experiment 3: Modelling Cross-Modal Congruence

TO DO The final experiment moves beyond direct regression to investigate the underlying relationship between the emotional content of the two modalities. A Siamese network is designed to explicitly model the semantic similarity between audio and lyrical representations.

3.6.1 Siamese Network Architecture for Similarity Learning

TO DO

3.6.2 Data Pairing and Contrastive Training Objective

TO DO

3.6.3 Evaluation Protocol for Similarity Score

TO DO

Chapter 4

Results

This chapter presents the empirical results obtained from the experiments detailed in the previous chapter. These are organised into sections covering the primary experimental outcomes using the MERGE dataset, a comparative analysis on a second dataset, and a final similarity analysis.

TO DO - ADD IN NAME OF SECOND DATASET HERE AND ADD IN ANYTHING MORE ABOUT THE SIMILARITY. remmove anything in red if I dont do a secon dataset

4.1 Experiment 1 Results: MERGE Dataset

TO DO REWRITE THIS SECTION BASED ON NEW RESULTS, EACH PARAGRAAPH NEEDS REWRITIN, DONE BLOCK 1 NEEED TO CHECK AGAIN AND ENSURE ALL REFERENCES ARE CORRECT.

The lightweight VGG-style model(Model 1) produced a surprisingly competitive performance, particularly the R^2 score for Valence, which exceeded the results from several models cited in the literature Jiang et al. (2024). Furthermore, Model 1 model produced unexpectedly stronger results for the R^2 for Valence than Arousal, a counter-intuitive result as, according to the literature, Valence is generally considered more difficult to predict Yang and Chen (2012).

The second model, which incorporated the batch normalisation, optimisation, and regularisation techniques detailed in the previous chapter, yielded a significant performance improvement in the R^2 score, particularly for Arousal, compared to the initial model. However, the third model, which incorporated the feature-wise attention mechanism, did not produce an improvement and all metrics were slightly decreased compared to the second model.

Finally, the fourth model, which implemented a true VGG-style architecture with dual convolutions per block, showed similar results to the second model, whilst requiring higher computational resources. This is an interesting result as it shows that architectural complexity does not necessarily translate to better performance.

In contrast to frozen BERT approaches, this model allowed all BERT parameters to be fine-tuned during training. However, this resulted in degraded performance compared to simpler models, with notable decreases in both Valence and Arousal R^2 scores. The model exhibited signs of overfitting, with training loss continuing to decrease whilst validation performance plateaued after epoch 4, ultimately requiring early stopping after 29 epochs due to lack of improvement.

Lastly, the fifth model, which used BERT without the frozen weights resulted in degraded performance compared to simpler models. The performance of all five models is shown in Table 4.1.

Table 4.1: Performance comparison of the four CNN model configurations for Valence and Arousal prediction. Arrows indicate the preferred direction for each metric.

Model Configuration	Dimension	MSE ↓	RMSE ↓	MAE ↓	R² ↑
1. Lightweight VGG-Style CNN	Valence	0.0240	0.1549	0.1185	0.5096
	Arousal	0.0088	0.0938	0.0722	0.3811
2. Model 1 + Batch Norm & Regularisation	Valence	0.0234	0.1530	0.1157	0.5219
	Arousal	0.0066	0.0812	0.0602	0.5330
3. Model 2 with Refined Feature-wise Attention	Valence	0.0255	0.1597	0.1219	0.4780
	Arousal	0.0066	0.0812	0.0593	0.5333
4. True VGG-style Architecture (2 Convs per Block)	Valence	0.0221	0.1487	0.1112	0.5481
	Arousal	0.0065	0.0806	0.0590	0.5420
5. Model 3 + Unfrozen BERT	Valence	0.0223	0.1493	0.1018	0.5444
	Arousal	0.0091	0.0954	0.0722	0.3553

Due to the surprisingly positive results, especially the error metrics that appeared to outperform many state-of-the-art benchmarks, an investigation into the dataset construction was performed. This was essential as no other research had been published using this new dataset. The investigation highlighted that whilst Louro et al. made the Valence and Arousal values available so MERGE was suitable for regression, they also filtered out ambiguous songs. They achieved this by removing any songs for which the average value of Valence or Arousal was in the range of -0.2 and 0.2 (corresponding to a 0.4 to 0.6 range on a 0-1 normalised scale), as illustrated in Figure 4.1. This had a profound effect on the suitability of this dataset for regression tasks, as the model could simply learn to avoid the empty central region and would be essentially performing a classification task, disguised as regression.

To check if the results were due to the data distribution, a series of benchmark tests were constructed. These confirm what can be observed visually; the data points are tightly clustered into distinct groups away from decision boundaries, which may incentivise a model to predict the central point of the clusters.

The following baseline models were constructed:

- **Mean Baseline:** This naive model predicts the overall mean Valence and Arousal value for every song.
- **Random Baseline:** This model predicts a random value between 0 and 1 for both Valence and Arousal, which represents a model with no learnt knowledge.
- **Quadrant-Centre Baseline:** This model has access to the quadrant classification data from the test data and predicts the mean of that quadrant as its output.

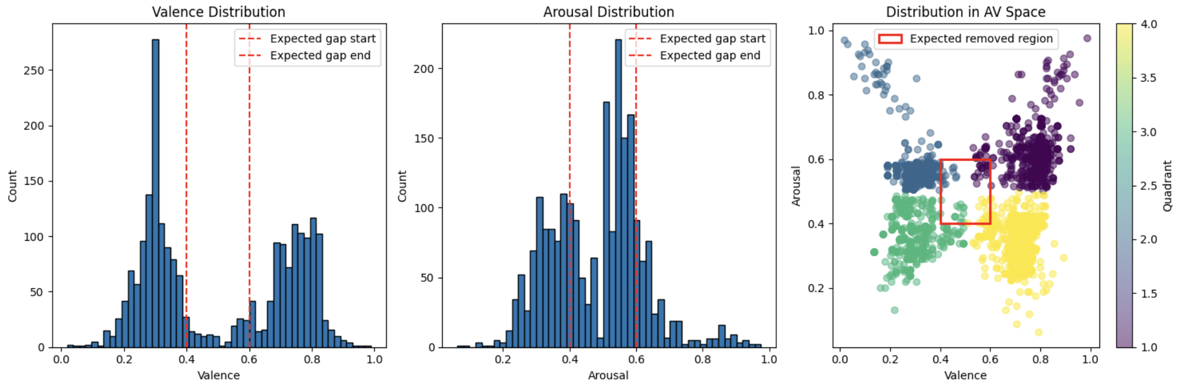

Figure 4.1: Visualisation of the MERGE dataset's data distribution. The histograms for Valence and Arousal show a central gap in the 0.4 - 0.6 normalised range, and the scatter plot shows the empty central region of the plot where the ambiguous songs were removed.

The results in Table 4.2 indicate that the model has learnt something meaningful, as it outperforms the first two baseline models. However, it has not performed as well as the Quadrant-Centre Baseline. This finding substantiates the claim that the most effective strategy for the CNN model is to perform a psuedo-regression task, always predicting the mean of the cluster, which is effectively a classification task. Although this wasn't the primary intention of the experiment, it still demonstrates meaningful learning. The model must distinguish between the emotional quadrants and predict appropriate continuous values relative to the central point of each cluster.

Table 4.2: Mean Absolute Error (MAE) comparison between the best CNN model and baselines.

Model	**Valence MAE ↓**	**Arousal MAE ↓**
Random Baseline	0.3245	0.2750
Mean Baseline	0.2088	0.1062
CNN Model Configuration 2	**0.1146**	**0.0617**
Quadrant-Centre Baseline	**0.0490**	**0.0423**

4.2 Experiment 2 Results: [Second Dataset Name]

The model was retrained and evaluated on the [Second Dataset Name] using an identical protocol to allow for direct comparison.

4.2.1 Quantitative Regression Performance

4.2.2 Visualisation of Predicted vs. True Values

4.3 Experiment 3 Results: Cross-Modal Congruence

The Siamese network was trained to produce a similarity score between audio and lyric pairs. The primary result of this experiment is the analysis of the relationship between this learned similarity and the prediction error of the regression model.

4.3.1 Learned Similarity Distribution

4.3.2 Correlation Between Modality Congruence and Prediction Error

Chapter 5

Analysis of Results

Your analysis needs a clear reference to the general goals of your proposal, and the objectives that you have reached in demonstrating your point.

You need to give reasoned explanations of your results and you also need to justify the way you have analysed and evaluated your results.

Unexpected or negative or limited results should also be reported and analysed. Why did they occur? What could have been done differently? Pitfalls happen, but make sure that you are adding a critical discussion of your results no matter how they are

Give an honest and balanced discussion, specific and not general.

What your investigation has determined, based on the results obtained.

How significant are the results you have obtained in your investigation? What are the significant factors and why are they significant?

Whether the experiments you chose were appropriate, based on the usefulness of the experimental results you obtained;

If the experiments did not provide useful results, or mediocre results, how they could be redesigned to be more useful and appropriate;

Whether your methodology for the investigation was appropriate. Why or why not? Howwould you change your overall methodology if you had to do the project again, to obtain more useful results?

This chapter provides an in-depth analysis and interpretation of the results presented in the previous chapter. It moves beyond the raw numbers to discuss the implications of the findings, address the research questions, and synthesise the outcomes of the three experiments into a cohesive narrative.

5.1 Analysis of Baseline Performance on the MERGE Dataset

The results from Experiment 1 revealed a distinct and informative pattern in the model's predictive behaviour, which forms a critical point of analysis for this dissertation.

5.1.1 Interpreting the Quantitative Metrics

5.1.2 The Central Clustering Phenomenon: Analysing the 'Doughnut'

5.1.3 Implications of MSE Loss on a Bounded Distribution

5.2 Comparative Analysis: MERGE vs [Second Dataset Name]

By comparing the outcomes of Experiment 1 and Experiment 2, we can draw conclusions about the model's robustness and the influence of dataset-specific properties.

5.2.1 Comparing Performance Metrics Across Datasets

5.2.2 Persistence of the Central Clustering Effect

5.3 Discussion of Cross-Modal Congruence and its Impact

The findings from Experiment 3 provide a potential explanation for the challenges observed in the regression tasks. This section discusses the critical relationship between how well the modalities agree and how well the model can predict the emotion.

5.3.1 Validating the Hypothesis: Is High Congruence Linked to Low Error?

5.3.2 Implications for Multimodal MER

5.4 Chapter Summary and Synthesis

This section synthesises the key analytical points from all three experiments, summarising what has been learned and how it contributes to the field of Music Emotion Recognition.

Chapter 6

Conclusions

This is the chapter in which you review the major achievements in the light of your original objectives, critique the process, critique your own learning and identify possible future work.

A critical evaluation of your work and a discussion of the conclusions that can be drawn from it is an essential part of your dissertation.

it should be a critical review and evaluation of your project. We would like to hear your own voice as the "owner" of your project!

First, state what you have achieved in the project (contributions of your work). State which of the original requirements have been fulfilled, and which have not (if any).

Indicate any extra work you have done, above and beyond the original requirements. If your requirements had to change as you went through the project, describe why they had to change and how they changed.

Evaluate what you have achieved. Is it what you expected (if not, why not)? Is it good, bad, or indifferent?

How significant are your results - what are the significant factors and why are they significant?

Do your achievements tell us anything new about the subject area you were working in?

Would you approach the problem differently if you had to do it again? Why?

Was your overall methodology for the project appropriate? Why or why not?

How would you change your overall methodology if you had to do the project again?

Make suggestions for future work which needs to be done.

Has your project suggested any new or unusual directions that should be considered? NOTES In the merge paper for merge they mention the construction - The values were mapped from using the all music tags and using a dictionary warner. This has been averaged so is this trustable? would like ti know more about this and re-evaluate the ratings. they did have 8 people that went over and re-evaluated, but what about the 8 peoples preferences. its hard to find emotion in some types of music, for example rock music. you have to enjoy rock music to hear the differences in a metal song.

Also need to look into whether the whole lyrics were included rather thajust the lyrics for the section of the song that were included. this could change the output as the song may change and the lyrics change with it later in the song. if this hasn't been done then we need to suggest this as an improvement. they are all included - for example 1111 is just guitar intro with about 5 words. the whole song lyrics are incuded. this might help or hinder the model idf the song is the same emotion all the way through then it might give the model a clue on what the song holds, but if the song is sa story and changes, then this could affect the model.

i used max pooling but it would be interesting to know whether the max pooling would be better with something else - looking at the Zhao paper that I have T-Max-Avg Pooling Layer: might be an interesting future question.

discuss Fong paper where they discuss that cnns are deisgned for image and this might not be the best for a spectrogram NNs are deep neural networks specially developed for image processing, in which objects are contiguous across both x- and y-dimensions, which have spatial meaning based on physical reality. Convolution filters play a critical role in determining the performance of CNNs, and the filters designed for image processing take advantage of spatial contiguity to perform effectively. However, spectrograms generated from music audio are not like reg- ular images. In spectrograms, non-contiguous regions in the frequency space impact many characteristics of music that humans perceive. For example, the simultaneous playing of an octave (e.g., A4 (440 Hz) and A5 (880 Hz)) produces a consonant sound while the simulta- neous playing of a tritone (e.g., A4 (440 Hz) and D5 (587 Hz)) produces a dissonant sound. Typical square CNN filters, which account for contiguous areas of an image, cannot cap- ture such concepts based on non-contiguous frequencies, thus highlighting the challenge of designing novel filters to incorporate music domain knowledge into deep learning models. 9 Goodfellow et al. (2016) write in their textbook the following about square convolution filters: "When a task involves incorporating information from very distant locations in the input, then the prior imposed by convolution may be inappropriate."

Bibliography

Cazzaniga, S., Gasparini, F. and Saibene, A., 2024. A multi-source deep learning model for music emotion recognition. *Proceedings of the 3rd Workshop on Artificial Intelligence for Human-Machine Interaction 2024 co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024)*. Bolzano, Italy: CEUR-WS, vol. 3903, pp.33–43.

Chaturvedi, V., Kaur, A.B., Varshney, V., Garg, A., Chhabra, G.S. and Kumar, M., 2022. Music mood and human emotion recognition based on physiological signals: a systematic review. *Multimedia systems* [Online], 28(1), pp.21–44. Available from: `https://doi.org/10.1007/s00530-021-00786-6` [Accessed 2025-10-14].

Fong, H., Kumar, V. and Sudhir, K., 2025. A Theory-Based Explainable Deep Learning Architecture for Music Emotion. *Marketing science* [Online], 44(1), pp.196–219. Available from: `https://doi.org/10.1287/mksc.2022.0323` [Accessed 2025-10-14].

Han, D., Kong, Y., Han, J. and Wang, G., 2022. A survey of music emotion recognition. *Frontiers of computer science* [Online], 16(6), p.166335. Available from: `https://doi.org/10.1007/s11704-021-0569-4` [Accessed 2025-10-21].

Han, X., Chen, F. and Ban, J., 2023. Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure. *Electronics* [Online], 12(4), p.978. Available from: `https://doi.org/10.3390/electronics12040978` [Accessed 2025-10-21].

Jiang, X., Zhang, Y., Lin, G. and Yu, L., 2024. Music Emotion Recognition Based on Deep Learning: A Review. *Ieee access* [Online], 12, pp.157716–157745. Available from: `https://doi.org/10.1109/ACCESS.2024.3484470` [Accessed 2025-10-05].

Liyanarachchi, R., Joshi, A. and Meijering, E., 2025. A Survey on Multimodal Music Emotion Recognition [Online]. ArXiv:2504.18799 [cs]. Available from: `https://doi.org/10.48550/arXiv.2504.18799` [Accessed 2025-07-22].

Louro, P.L., Redinho, H., Malheiro, R., Paiva, R.P. and Panda, R., 2024. A Comparison Study of Deep Learning Methodologies for Music Emotion Recognition. *Sensors* [Online], 24(7), p.2201. Available from: `https://doi.org/10.3390/s24072201` [Accessed 2025-11-04].

Louro, P.L., Redinho, H., Santos, R., Malheiro, R., Panda, R. and Paiva, R.P., 2025. MERGE – A Bimodal Audio-Lyrics Dataset for Static Music Emotion Recognition [Online]. ArXiv:2407.06060 [cs]. Available from: `https://doi.org/10.48550/arXiv.2407.06060` [Accessed 2025-09-30].

Ma, Y., Øland, A., Ragni, A., Sette, B.M.D., Saitis, C., Donahue, C., Lin, C., Plachouras, C., Benetos, E., Shatri, E., Morreale, F., Zhang, G., Fazekas, G., Xia, G., Zhang, H., Manco, I., Huang, J., Guinot, J., Lin, L., Marinelli, L., Lam, M.W.Y., Sharma, M., Kong, Q., Dannenberg, R.B., Yuan, R., Wu, S., Wu, S.L., Dai, S., Lei, S., Kang, S., Dixon, S., Chen, W., Huang, W., Du, X., Qu, X., Tan, X., Li, Y., Tian, Z., Wu, Z., Wu, Z., Ma, Z. and Wang, Z., 2024. Foundation Models for Music: A Survey [Online]. ArXiv:2408.14340 [cs]. Available from: `https://doi.org/10.48550/arXiv.2408.14340` [Accessed 2025-05-27].

Pandeya, Y.R. and Lee, J., 2024. GlocalEmoNet: An optimized neural network for music emotion classification and segmentation using timbre and chroma features. *Multimedia tools and applications* [Online], 83(30), pp.74141–74158. Available from: `https://doi.org/10.1007/s11042-024-18246-4` [Accessed 2025-10-14].

Pyrovolakis, K., Tzouveli, P. and Stamou, G., 2022. Multi-Modal Song Mood Detection with Deep Learning. *Sensors* [Online], 22(3), p.1065. Available from: `https://doi.org/10.3390/s22031065` [Accessed 2025-10-14].

Qiao, Y., Mu, J., Xie, J., Hu, B. and Liu, G., 2024. Music emotion recognition based on temporal convolutional attention network using EEG. *Frontiers in human neuroscience* [Online], 18, p.1324897. Available from: `https://doi.org/10.3389/fnhum.2024.1324897` [Accessed 2025-10-21].

Su, Y., Chen, J., Chai, R., Wu, X. and Zhang, Y., 2024. FFA-BiGRU: Attention-Based Spatial-Temporal Feature Extraction Model for Music Emotion Classification. *Applied sciences* [Online], 14(16), p.6866. Available from: `https://doi.org/10.3390/app14166866` [Accessed 2025-10-21].

Wang, J., Sharifi, A., Gadekallu, T.R. and Shankar, A., 2024. MMD-MII Model: A Multilayered Analysis and Multimodal Integration Interaction Approach Revolutionizing Music Emotion Classification. *International journal of computational intelligence systems* [Online], 17(1), p.99. Available from: `https://doi.org/10.1007/s44196-024-00489-6` [Accessed 2025-10-14].

Won, M., Spijkervet, J. and Choi, K., 2021. Music Classification: Beyond Supervised Learning, Towards Real-world Applications [Online]. [Online], abs/2111.11636. Available from: `https://doi.org/10.5281/ZENODO.5703779`.

Yang, L., Shen, Z., Zeng, J., Luo, X. and Lin, H., 2023. COSMIC: Music emotion recognition combining structure analysis and modal interaction. *Multimedia tools and applications* [Online], 83(5), pp.12519–12534. Available from: `https://doi.org/10.1007/s11042-023-15376-z` [Accessed 2025-11-04].

Yang, Q., Liu, S. and Gong, T., 2025. Improve the application of reinforcement learning and multi-modal information in music sentiment analysis. *Expert systems* [Online], 42(1), p.e13416. Available from: `https://doi.org/10.1111/exsy.13416` [Accessed 2025-10-21].

Yang, Y.H. and Chen, H.H., 2012. Machine Recognition of Music Emotion: A Review. *Acm transactions on intelligent systems and technology* [Online], 3(3), pp.1–30. Available from: `https://doi.org/10.1145/2168752.2168754` [Accessed 2025-11-11].

Yazhong Feng, Yueting Zhuang and Yunhe Pan, 2003. Music information retrieval by detecting mood via computational media aesthetics [Online]. *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*. Halifax, NS, Canada: IEEE Comput. Soc, pp.235–241. Available from: https://doi.org/10.1109/WI.2003.1241199 [Accessed 2025-11-04].

Zhao, L. and Zhang, Z., 2024. A improved pooling method for convolutional neural networks. *Scientific reports* [Online], 14(1), p.1589. Available from: https://doi.org/10.1038/s41598-024-51258-6 [Accessed 2025-10-21].

Appendix A

Design Diagrams

Appendix B

User Documentation

Appendix C

Raw Results Output

Appendix D

Code

D.1 File: yourCodeFile.java

```java
// This is an example java code file, just for illustration purposes
public static void main() {
    System.out.print ("Hello␣World");
}
```