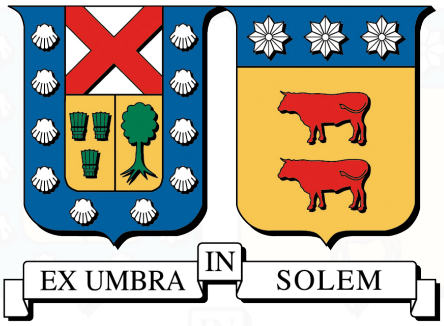
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
ELECTRONICS ENGINEERING DEPARTMENT
VALPARAÍSO - CHILE



A Transformer Approach for the Analysis of Music's Emotional Trajectory Directly from Music Audio for Recommender System Applications

Submitted by

Pascal Arriagada Stoller

In partial fulfillment of the requirements for the award of the degree of

MASTER OF SCIENCE
IN ELECTRONICS ENGINEERING

and title of

TELEMATICS ENGINEER

THESIS ADVISOR	:	Ph.D. MAURICIO ARAYA LÓPEZ.
INTERNAL REVIEWER	:	Ph.D. MATÍAS ZAÑARTU SALAS
EVALUATION COMMITTEE	:	Ph.D. SEBASTIÁN MORENO ARAYA

MARCH 2024

ABSTRACT

Under the research field of Music Dynamic Emotion Recognition, a subset of Music Information Retrieval and Affective Computing, this study delves into research on Deep Learning techniques focused on the estimation of values associated with the perceived *emotional trajectory* of complete musical pieces or excerpts. Based on this, one of the research objectives consisted of the implementation of a transformer model that provides dynamic predictions, which are then used for an emotion-based Musical Recommendation System. The concept of emotional trajectory is defined under Russell's Circumplex model of Affect (1980), a dynamic emotional theory based on the orthogonal dimensions of Arousal and Valence. The models to implement utilize as input a set of 260 low-level features extracted via the openSMILE toolkit. These features were selected during the generation process of the MediaEval dataset for Emotional Analysis in Music (2015), which comprises 1802 music pieces along with dynamic emotional annotations. This dataset serves as a benchmark in the field, offering selected metrics to assess model performance and predictive behavior, obtaining results similar to the state-of-the-art, thus demonstrating the applicability of this type of user-oriented systems.

RESUMEN

Bajo el area de estudio del Reconocimiento Dinámico de Emociones en la Música, subárea del campo de Recuperación de Información Musical y la Computación Afectiva, esta investigación se enfoca en técnicas de Deep Learning orientadas a la estimación de valores asociados a la *trayectoria emocional percibida* de piezas musicales completas o fragmentos de estas. Para ello, se implementó un modelo *Transformer* que obtiene predicciones dinámicas, las cuales son usadas como base para un Sistema de Recomendación Musical basado en emociones. El concepto de *trayectoria emocional* se define bajo la teoría emocional dimensional de Russell (1980), quien presenta un modelo circunflejo de emociones usando las dimensiones ortogonales de *excitación cortical y alerta* (Arousal) y *valencia* (Valence). Los modelos implementados se construyen sobre un conjunto de 260 características de bajo nivel obtenidas con el toolkit openSMILE, las cuales fueron seleccionadas en el proceso de generado del dataset MediaEval para Análisis Emocional en la Música (2015), el cual consiste en 1802 canciones disponibles junto con sus respectivas anotaciones emocionales estáticas y dinámicas. Este dataset fue diseñado para ser utilizado como *benchmark* en el área, por lo cual cuenta con métricas seleccionadas para validar el rendimiento y comportamiento de los modelos desarrollados, obteniendo resultados similares al estado del arte, demostrando así la aplicabilidad de este tipo de sistemas enfocados en el usuario.

Contents

Abstract i

Resumen ii

List of Tables v

List of Figures vii

1 Introduction 1
- 1.1 Motivation and Scope . . . 2
- 1.2 Hypothesis . . . 3
- 1.3 Objectives . . . 4
- 1.4 Thesis Overview . . . 5

2 Area of study 7
- 2.1 Affective Computing . . . 8
  - 2.1.1 Music Information Retrieval . . . 9
  - 2.1.2 Music Emotion Recognition . . . 10
  - 2.1.3 Music Recommendation Systems . . . 13
- 2.2 Emotion Taxonomies . . . 14
  - 2.2.1 Discrete . . . 15
    - 2.2.1.1 Ekman “basic emotions” . . . 16
    - 2.2.1.2 Hevner “Affective Ring” . . . 17
  - 2.2.2 Dimensional . . . 17
    - 2.2.2.1 Circumplex model of Affect . . . 19
    - 2.2.2.2 Thayer’s Mood model . . . 20
- 2.3 Emotional Trajectory . . . 20
- 2.4 Datasets . . . 22
  - 2.4.1 DEAM . . . 22
  - 2.4.2 MERP . . . 26
  - 2.4.3 MUSAV . . . 27
  - 2.4.4 AVEC . . . 28
- 2.5 Feature Engineering . . . 29

3 Literature review 31
- 3.1 Prediction Approaches for MER systems . . . 31
  - 3.1.1 Classical Techniques . . . 31
  - 3.1.2 Convolutional Models . . . 32
  - 3.1.3 Recurrent Networks . . . 33
  - 3.1.4 Transformer . . . 35
    - 3.1.4.1 Model Overview . . . 35
    - 3.1.4.2 Highlighted Works Reviewed . . . 37
- 3.2 Prediction Approaches of Emotional Trajectory . . . 40
- 3.3 Benchmark . . . 40

3.3.1 DEAM 40

4 Emotional Dynamics Models **43**
4.1 Data Preprocessing 44
4.2 MLP Baseline 46
4.2.1 Model Definition 46
4.2.2 Implementation 48
4.2.2.1 Static 49
4.2.2.2 Dynamic averaged 51
4.2.2.3 Dynamic 53
4.2.3 Model Analysis 55
4.2.3.1 Top CCC models 56
4.2.3.2 Metrics comparison 57
4.3 Transformer Model Configurations 61
4.3.1 Model Definition 61
4.3.2 Implementation 63
4.3.2.1 Unimodular 64
4.3.2.2 Bimodular 66
4.3.3 Model Analysis 68

5 Emotional System **72**
5.1 Emotional Trajectory Estimation 73
5.1.1 Reference values 73
5.1.2 Estimated values 76
5.1.3 Estimation Analysis 78
5.2 Recommender System 81

6 Conclusions **89**
6.1 Summary and Conclusions 89
6.2 Perspectives for Future Research 92
Bibliography 93

A Transformer Model Code Snippet **101**

List of Tables

2.1 OpenSMILE's low-level descriptors [1]. 30

3.1 RMSE metrics for Arousal and Valence dimensions. Models available in the working notes papers of MediaEval 2015 Emotion in Music Task [2]. 41

4.1 Top 5 models with lowest RMSE scores implemented over the static dataframe testing set. . 49
4.2 Top 5 models with lowest RMSE scores implemented over the static dataframe evaluation set. 49
4.3 Top models implemented over the evaluation set of the static dataframe. 50
4.4 Top 5 models with lowest RMSE scores implemented over the dynamic averaged dataframe testing set. 51
4.5 Top 5 models with lowest RMSE scores implemented over the dynamic averaged dataframe evaluation set. 51
4.6 Top models implemented over the evaluation set of the dynamic averaged dataframe. 52
4.7 Top 5 models with lowest RMSE scores implemented over the dynamic dataframe testing set. 53
4.8 Top 5 models with lowest RMSE scores implemented over the dynamic dataframe evaluation set. 53
4.9 Top models implemented over the evaluation set of the dynamic dataframe. 54
4.10 Top 5 models with highest CCC scores implemented over the dynamic dataframe testing set. 56
4.11 Top 5 models with highest CCC scores implemented over the dynamic dataframe evaluation set. 56
4.12 Top models implemented over the evaluation set of the dynamic dataframe. 56
4.13 Reported *best test scenarios* of Medina *et al.* work. 58
4.14 Top 5 models with the lowest RMSE scores implemented in the Unimodular approach, comparing cases with and without the LSTM layer. 64
4.15 Top 5 models with the highest CCC scores implemented in the Unimodular approach, comparing cases with and without the LSTM layer. 64
4.16 Top 5 models with the lowest RMSE scores implemented in the Bimodular approach, comparing cases with and without the LSTM layer. 66
4.17 Top 5 models with the highest CCC scores implemented in the Bimodular approach, comparing cases with and without the LSTM layer. 66
4.18 Top models with the best scores implemented in Unimodular and Bimodular approaches. . . 69

5.1 Time distribution comparison between the estimated emotional trajectory values and the reference values for Song # 2012. 87
5.2 Time distribution comparison between the estimated emotional trajectory values and the reference values for Song # 2027. 87
5.3 Top 5 emotion percentage distribution for Song # 2012. 88
5.4 Top 5 emotion percentage distribution for Song # 2027. 88

List of Figures

2.1 Hevner affective ring model [3]. . . . 18
2.2 Two-dimensional circumplex space model. This emotional plot was adapted from a list of emotions and moods and their correspondent V/A values available in the works of Paltoglou & Thelwall [4] and Coutinho et al. [5]. . . . 19
2.3 Thayer's model. Energy refers to the volume or intensity of sound in music, and Stress refers to the tonality and tempo of music. The mood is divided into four clusters: calm-energy (Exuberance), calm-tiredness (Contentment), tense-energy (Frantic or Anxious), and tense-tiredness (Depression) [6]. . . . 20
2.4 Web annotation interface used for registering dynamic and static arousal and valence ratings for each song after it has been listened to at least once [7]. . . . 23
2.5 DEAM V/A values distribution per song for each approach: static, dynamic avg., and dynamic. 24
2.6 (a) Representation of the predicted arousal and valence value of the top 1000 songs of the FMA (filtered by length criteria). Colored dots represent the songs selected for their investigation. (b) Annotation interface on which arousal and valence values were captured via mouse tracking in the listening study [8]. . . . 26
2.7 Presented baseline architectures [8]. **Left**: Fully Connected architecture. **Right**: LSTM architecture. . . . 27
2.8 Web annotation interface employed for recording comparative arousal and valence ratings for each song. Figure adapted from [9]. . . . 28

3.1 Transformer model architecture presented by Vaswani *et al.* [10]. . . . 36
3.2 (left) Scaled Dot-Product Attention. (right) Multi-head attention consists of several attention layers running in parallel. Vaswani *et al.* [10]. . . . 36

4.1 Flowchart depicting the implementation of the typical MER system steps in this work and the proposed Emotional System. . . . 43
4.2 The flowchart details the steps of the Data Preprocessing Process using the values from the DEAM dataset. . . . 44
4.3 DEAM V/A dynamic values distribution per timestamp with RGB color model based in Dharmapriya *et al.* work [11]. Red and orange colors represent high arousal values, and blue and green represent low arousal values, a convention also used by Coutinho *et al.* [12]. . . . 46
4.4 The principal phases of the Emotion Prediction System presented by Medina *et al.* [13]. . . . 47
4.5 Visualization of the predictions obtained with the Top model for DEAM static dataframe approach. . . . 50
4.6 Visualization of the predictions obtained with the Top model for DEAM dynamic averaged dataframe approach. . . . 52
4.7 Visualization of the predictions obtained with the Top model for DEAM dynamic dataframe approach. . . . 54
4.8 Visualization of the predictions obtained with top CCC models for DEAM dynamic dataframe approach. . . . 57
4.9 A Violin Plot comparison depicting the variation of RMSE, MAE, and CCC metrics for the Valence dimension based on top metric selection criteria. . . . 59
4.10 A Violin Plot comparison depicting the variation of RMSE, MAE, and CCC metrics for the Arousal dimension based on top metric selection criteria. . . . 60

4.11 In the left (4.11a), the multimodal Transformer model presented in by Huang *et al*. [14]. (4.11b, 4.11c) Transformer models proposed for MEVD analysis directly from music audio features by deconstructing the model. 61
4.12 Unimodular model architecture with LSTM. 65
4.13 Visualization of the predictions obtained with the unimodular approach for top RMSE and CCC models. 65
4.14 Bimodular model architecture with LSTM. 67
4.15 Visualization of the predictions obtained with the Bimodular approach for top RMSE and CCC models. 68
4.16 A Violin Plot comparison depicting the variation of RMSE, MAE, and CCC metrics for the Unimodular approach dimension based on top metric selection criteria. 70
4.17 A Violin Plot comparison depicting the variation of RMSE, MAE, and CCC metrics for the Bimodular approach dimension based on top metric selection criteria. 71

5.1 Referential Emotional Trajectory Visualization for Song # 2012. 74
5.2 Referential Emotional Trajectory Visualization for Song # 2027. 75
5.3 Estimated Emotional Trajectory Visualization for Song # 2012. 76
5.4 Estimated Emotional Trajectory Visualization for Song # 2027. 77
5.5 Windowed CCC metric values visualization for Songs # 2012 and 2027. 80
5.6 Emotion distribution and class distribution of the emotional tags based on the Russell circumplex model. Figure drawn from [15]. 82
5.7 Valence and Arousal Reference values discretized to High and Low categories. Visualization for Song # 2012. 83
5.8 Valence and Arousal Reference values discretized to High and Low categories. Visualization for Song # 2027. 84
5.9 Valence and Arousal Estimated values discretized to High and Low categories. Visualization for Song # 2012. 85
5.10 Valence and Arousal Estimated values discretized to High and Low categories. Visualization for Song # 2027. 86
5.11 Estimated distribution of the emotional tags presented in Subsubsection 2.2.2.1 over each dimension. 88

1 | Introduction

Listening to music plays a significant role in our daily lives, influencing our emotions and overall mood [16]. The music we choose, the emotions it evokes, the ones we identify, and the resulting mood are all interrelated in ways we are unbeknownst to. Music is perhaps the most socially accepted means of mood manipulation, with individuals often turning to music to uplift or calm themselves [17]. It is common for people to choose music that resonates with their current emotional state [18] or deliberately select music to evoke a desired emotional response. Nowadays, it is practically impossible to predict with exactitude which piece of music somebody would most like to listen to at any moment, although advances in EEG research should make this possible in a few years [19–21]. On the other hand, recommending a general type or genre of music that someone would prefer at a given time is a more approachable challenge [22, 23]. People usually like to have a choice but do not like being overwhelmed with too many choices, which constrains how precise or variable the recommendation must be. It is important to consider that in addition to musical taste, mood is a big influence on musical choice, as many people make simple selections based on how calm or wired they feel.

Furthermore, the proliferation of available music resources has led to research into organizing and labeling musical pieces with emotional values to enhance user recommendations. However, manual methods for obtaining emotion labels are time-consuming, labor-intensive, and error-prone, which led to the emergence of the research field of automatic emotion recognition, the backbone of the present investigation.

1.1 Motivation and Scope

This research is driven by a deep interest in obtaining personalized music recommendations that align with specific emotional labels or sets of emotional labels, subsequently referred to as an *emotional trajectory*. As a result, the concept of an Emotional System for Dynamic Music Emotion Recognition emerged as the foundational idea behind this study. To achieve this goal, a thorough review of the associated research area is essential to understand better the context, limitations, and evaluation procedures of implementations in this field. The ideal system is user-driven, prioritizing fast and dynamic access to the system in a lightweight and versatile way. However, this constrains the problem by limiting the useful musical information in the pieces.

The definition of an Emotional System's main steps and subsequent implementation workflow is an interesting challenge, as it comprises selecting the Emotional Taxonomy alongside a dataset and related benchmarks on which to base the implementation and validation of the prediction models. Moreover, the objective of using this emotional system as an emotion-based music recommendation system adds another layer to the research: the definition of recommendation criteria.

On the technical side, among the main techniques and model implementations reviewed for Music Emotion Variation Detection (MEVD) and Music Emotion Recognition (MER) systems, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)-based approaches dominate the field. Andayani *et al.* [24] stated that RNNs have been widely used in MEVD research due to their ability to process sequential data and handle variable-length input. However, they suffer from the long-term dependencies problem, and cannot utilize parallelization in their architecture. On the other hand, LSTM models can capture long-term dependencies using their memory architecture, enabling them to retain information over extended periods. However, parallelization remains challenging for LSTM and might not adequately handle longer-term dependencies.

The Transformer method, introduced in 2017, employs a Multi-Head Attention Mechanism and has found widespread use in various fields, particularly in Natural Language Processing (NLP). This method helps overcome the drawbacks of parallelization in the

RNN/LSTM algorithm. Additionally, Multi-Head Attention in the Transformer allows it to simultaneously process information from different feature subspaces at different sequence positions, enhancing its ability to learn temporal information. This characteristic has been recently exploited in different emotion recognition approaches as more researchers are implementing Transformer-based systems.

1.2 Hypothesis

The perceived emotional trajectory of a piece of music by a person is feasible to estimate based on its low-level descriptors, with sufficient accuracy for applications of Emotion-based Music Recommendation Systems using deep-learning-based models trained on datasets annotated with empirically obtained Valence-Arousal (V/A) values.

Throughout the investigation, the above hypothesis was elaborated into more specific research questions addressing the design and implementation steps of the target emotional system:

1. *Implementing a Transformer-based model could improve the accuracy of MEVD MER systems.*

   Transformer models are useful in different research fields, particularly on NLP problems. They have also been applied to multimodal approaches of music emotion recognition, in which lyrics play a fundamental role. These approaches aim to predict or estimate the emotional behavior of the musical piece based on the content of the available lyrics. However, their potential in unimodal approaches, particularly those based solely on the low-level features of the pieces, still has associated research potential.

2. *The definition and application of a different metric could improve model performance evaluation, parameter selection and comparison, thereby improving the final system.*

   In the literature, researchers often use a particular set of metrics to evaluate how accurately models estimate and predict emotions. However, these metrics mainly focus on analyzing the prediction values at the song level, which might not fully

capture the dynamic emotional changes within the piece. A different metric approach, which considers the dynamic emotional prediction throughout the piece, is proposed to enhance the criteria for selecting models and improve prediction accuracy for quality recommendations.

3. *Applying the Emotional Trajectory concept in an Emotional System could enhance its recommendation capabilities.*

   The concept relates to emotional behavior in a piece of music. However, this value is not absolute; it is inherently subjective and depends on the annotations used to construct it. The more accurately this value represents the perceived emotional behavior in a piece of music, the more this representation can be used to describe that piece uniquely. Furthermore, exploiting the information in these values could improve the Emotional System's predictive capability and related accuracy.

1.3 Objectives

This study aims to design and implement an Emotion-based Music Recommendation System trained directly on music audio features. It incorporates the concept of *emotional trajectory* as the core of its recommendation process. The following objectives have been formulated:

1. Present the steps to implement the Music Emotion Recognition system based on state-of-the-art approaches.

2. To investigate existing models for MEVD and Transformer-based implementations in MER.

3. To implement the selected model architecture for the system.

4. Evaluate the performance of the models selected for the Emotional System using appropriate metrics and selected benchmarks.

5. Evaluate the impact of the proposed metric in the model selection step by comparing it with the models selected with state-of-the-art criteria.

6. Define the recommendation logic of the Emotional System in a user-driven manner based on the concept of Emotional Trajectory.

1.4 Thesis Overview

Chapter 1 Introduction

In this chapter, the overview of this research is presented. This chapter briefly discusses the research background, problem, and motivation. In addition, this chapter proposes the research questions, aim, and objectives of the identified research problem.

Chapter 2 Area of Study

This chapter covers a hierarchical analysis of the main concepts and context of the research topic. It explains the main emotional approaches and introduces the concept of an emotional trajectory linked to representing the dynamic emotional content of a piece of music. It outlines the chosen dataset and reviews other relevant datasets in the field. This chapter also discusses the evaluation metrics to be used in this study.

Chapter 3 Literature Review

This chapter presents a literature review focused on approaches taken in this context. It presents and describes the techniques and models generally used, gives a brief insight into different approaches that inspired this investigation, and provides a results comparison point. In parallel, the Transformer architecture is presented to compare the studied approaches based on this technique.

Chapter 4 Emotional Dynamics

This chapter discusses the implementation steps of a typical MER system adapted to the research objectives. Then, discusses and evaluates the performance results of the training and testing of the proposed models. This performance is evaluated based on the evaluation metrics presented in Chapter 2. Furthermore, the results are compared to other works proposed by existing researchers in the field, presented in Chapter 3.

Chapter 5 Emotional System

This chapter outlines the evaluation process of the model implementation results, comparing them with the estimated values. It also introduces an Emotion-based Music Recommender System Application built upon the implemented model and showcases alternative discretization and classification approaches.

Chapter 6 Conclusions

This chapter summarizes the main steps undertaken and the challenges encountered throughout the research and implementation processes. It also draws conclusions based on the findings and proposes areas for future improvements and research endeavors.

2 | Area of study

This chapter presents, from top to bottom, the main concepts underlying the research. Affective Computing is the root research field that encompasses the objective of providing a computational system with emotional capabilities that allow the development of more "human" systems, which can interact with humans more emotionally and naturally by recognizing the emotions expressed by humans or by predicting the emotions that can be perceived in daily interaction with such systems. Music Information Retrieval (MIR), a multidisciplinary research field, focuses on developing interfaces that allow users to navigate through vast collections of music more efficiently. Affective systems developed under this research field are grouped in the sub-task of Music Emotion Recognition (MER), an area of research focused on predicting or estimating the emotional impact of music through the computational analysis of musical properties, mapping these musical features to a previously defined emotion space.

Music Recommendation Systems (MRS) are interfaces designed to provide users with recommendations based on the user's preferences. In these systems, researchers can exploit the implementation of emotional capabilities that allow them to "understand" the emotional state of the user and return a recommendation related to that state or that can help the user to consciously change their current emotional state, by listening to music on which it can perceive a specific emotional concept.

A particular emotional taxonomy must be selected to develop these systems, to visualize the emotional response as a specific category or as a dynamic value that can vary over time. Based on the premise that the emotions in music change dynamically, the concept of *Emotional Trajectory* is introduced as a dynamic approach of MER focused on

predicting and estimating the emotional variations over time for a given piece of music. Finally, the MediaEval Dataset for Emotional Analysis of Music (DEAM), a database consisting of royalty-free song excerpts, a set of curated and selected low-level features, and source crowded static and dynamic annotations is introduced, alongside the selected evaluation metrics used to validate different approaches of MRS systems developed over this dataset.

2.1 Affective Computing

The term *Affective Computing* (AC) was introduced in 1997 by Rosalind Picard in a book named after the concept [17], describing it as "computing that relates to, arises from, or deliberately influences emotions". Among its main objectives, Picard included "giving a computer the ability to recognize and express emotions, developing its ability to respond intelligently to human emotions, and enabling it to regulate and utilize its emotions". Nowadays, AC is also known as *Emotion AI*, a vast area of research of considerable practical and theoretical interest to numerous fields, including signal processing, machine learning, computational linguistics, computer vision, neuroscience, cognitive science, and social psychology, that is concerned with emotion detection through the use of artificial intelligence [25]. This type of computing aims to bridge the gap between humans and machines, eventually allowing them to interact with humans naturally and emotionally, since two fundamental characteristics concerning human emotions are considered: the ability of other people to recognize an individual's feelings and the ability of an individual to feel that other people comprehend its emotional condition [26]. Thus, an essential objective of this research field is to develop systems that allow computers to recognize human subjects' emotional or affective states to generate personalized responses, making human-computer interaction more efficient and natural. In this context, the expression "recognizing emotions", according to Picard [17], should be interpreted as "inferring an emotional state from observations of emotional expressions and behavior, and through reasoning about an emotion-generating situation".

Aranha *et al.* in their 2021 systematic review article of Affective Computing software applications [27] states that, in a hypothetical situation, knowing the user's emotional

state allows the development of software that considers this information to, for example, recommend the adequate content. Additionally, such interventions in software are crucial because they allow, among other factors, the software to adapt to the user, while generally, it is the user who must adapt the software. Thus, AC generates a personalization that enables the user to take better advantage of the software resources.

2.1.1 Music Information Retrieval

Music Information Retrieval (MIR) is a multidisciplinary field part of the Information Retrieval research field. It emerged as a research field dedicated to developing innovative technologies and interfaces to assist the user in finding music, information about music, or information in music [28]. An early definition of this field was presented in 2004 by Downie [29], describing it as a "multidisciplinary research endeavor that strives to develop innovative content-based searching schemes, novel interfaces, and evolving networked delivery mechanisms to make the world's vast store of music accessible to all". Another definition is presented by Serra *et al.* in their 2013 book *Roadmap for Music Information ReSearch* [30], describing it as a research field focused on processing digital data related to music, including gathering and organizing machine-readable musical data, developing data representations, and methodologies to process and understand that data. A most recent definition can be found in the 2021 work of Gomez-Cañón *et al.* [31], where they presented MIR as an "interdisciplinary research field focused on developing computational systems to help humans better understand music collections, by integrating concepts and methodologies from several disciplines including music theory, music psychology, neuroscience, signal processing, and machine learning techniques".

The rapid and continuous development of the music industry, particularly the ever-growing electronic music market, has fueled the development of MIR systems to assist the user in navigating and exploring these large music collections [28]. The fact that there are over 589 million paid subscribers [32] and over 100 million tracks on Spotify alone (by 2023 [33]) shows the importance and potential research opportunities that can be tackled

from different approaches based on the core challenge to be addressed. MIR is described as being driven by a set of these core challenges, such as extracting meaningful features, indexing music for efficient retrieval, and designing search and retrieval tools like music recommendation systems (MRS) and content-based user interfaces (e.g., for browsing large music collections), which are most likely to be addressed from a user point of view. Schedl *et al.* in their survey "Music Information Retrieval: Recent Development and Applications" [34] describe these challenges as:

- **Music Retrieval**: Intended to help users find music in large collections by a particular similarity criterion. These scenarios could be classified according to *specificity* (high specificity to identify a given audio signal and low to get statistically similar or categorically similar music pieces) and *granularity* or temporal scope (large granularity to retrieve complete music pieces or fragments and small granularity to locate specific time locations or fragments).

- **Music Recommendation**: Typically propose a list of music pieces based on modeling the user's musical preferences. These systems consider the following requirements for their development: *accuracy* (recommendations that match musical preferences), *diversity*, *transparency* (that the user can understand why the recommendation has been made), and *serendipity* (how surprising the recommendation is).

- **Music Playlist Generation**: Automatic music playlist generation can be regarded as highly related to music recommendation. It aims to create a list of results, such as music tracks or artists, to provide meaningful, enjoyable playlists for the listener. Here, the order in which the tracks are given could be important, differentiating this application from music recommendation and the reorganization of already known material.

2.1.2 Music Emotion Recognition

Music Emotion Recognition (MER) is a MIR sub-task that belongs to the interdisciplinary research field of music psychology, audio signal processing, and natural language processing (NLP). Coutinho *et al.* [12] defined MER as "an area of research focused on estimating

the emotional impact of music through the computational analysis of musical properties". It is also stated that MER's basic premise is that music communicates and induces similar emotional states in all listeners because the musical parameters (e.g., rhythm, melody, timbre) encode affective information that listeners implicitly decode. Based on this, and as stated by Han *et al.* in their MER survey article [35], MER can be described as a process that is constituted of the extraction and analysis of music features to form mapping relations between music features and a particular emotional space, thus making it possible to recognize the emotion that is expressed by music. These music features are often extracted directly from the audio signal, symbolic music scores, and lyrics texts. Yang *et al.* [36] describes two main categories to group these features: audio and symbol files. Audio files, which store sound content in a waveform, are called the *low-level* representation of music. In contrast, symbol files, which describe sound information in bytes, are called the *high-level* representation of music. Based on this, different modalities are mainly studied solely or combined for the development of MER systems, which include and are not limited to *visual* (e.g., facial expressions, body gestures, eye gaze), vocal (e.g., speech, prosody & intonation), physiological (e.g., EEG, ECG, EDA, HR, PPG), and textual features, depending on the application context. For a user-driven MER system, these modalities constrain the implementation capabilities of the application, as the user would require extra gear in the case of physiological features or to add an extra layer of interaction to obtain and process the visual features. In this research, obtaining predictions directly from audio features is the chosen approach to extend the model predicting capabilities to any music input, regardless of the language or the availability of the lyrics due to copyright issues.

Yang *et al.* [36] summarize Music Emotion Recognition as the process of using computing systems to extract and analyze music features, establish mapping relations between music features and emotion space, and recognize the emotion expressed by music. Based on this, three main steps can describe a typical MER method. The first step is the Domain Definition Step, where the target and extent of the problem are defined by choosing different formats of music records and emotion models; in this step, the datasets are selected. This step can also be found separated in the literature [31], as it is necessary to first define an emotional taxonomy before handling the dataset creation or fetching. Second, music

features are extracted in the Feature Extraction step, and the ground-truth values are defined. These ground-truth values correspond to the annotations of the music pieces by different subjects based on a particular emotion model. In contrast, the music features consist of information that can be extracted from the music. The set and conditions to choose these features depend on the focus of the research. Finally, machine learning methods are used in the Model Training Process Step to establish the mapping relations between the selected music features and the emotions. Thus, an automatic MER system can be implemented.

The feature extraction step comprises one of the main difficulties associated with these systems. To achieve music emotion-based classification and retrieval, it is necessary to label music works with emotions [37], for which human-labeled ground truth annotations must be obtained. Still, this task often lacks a singular, well-defined answer [38]. This is also known as the *subjectivity problem*, which stems from the fact that music perception is intrinsically subjective and is influenced by many individual factors, such as personality, cultural background, musical preferences, or education. In the context of this study, MER is presented as the construction of a calculation model based on music audio data to achieve automatic music emotional predictions based on the concept of emotional trajectory. The objective of this system is to obtain models that can handle the workload associated with emotional annotations of music works and ensure that the estimations present a minimum quality and are useful to the user.

Another problem related to the first two steps, Domain Definition and Feature Extraction, arises due to audio copyright restrictions, which implies that data sets used in various studies are seldom made public and reused in other studies, thus limiting a fair comparison ground between models and techniques in the context. Annotations are often obtained by crawling the tags from social music websites, such as `last.fm` or `allmusic.com`. Still, the audio is usually copyrighted in this case, which doesn't allow researchers to redistribute them. The music distributed for free under a license such as Creative Commons is generally less known and has fewer tags; therefore, it must be annotated before being used in any study. In these cases, annotating the music pieces with emotional labels is a burdensome task, as many annotators are needed for every item to obtain a representative measure for such a subjective task [7].

2.1.3 Music Recommendation Systems

Music Recommendation Systems (MRS) are based under Picard's premise [17] that music would perhaps be the most socially accepted form of mood manipulation, as people often turn to music to cheer themselves up or to calm themselves down. Picard argues that, given that people usually like to have a choice but not to be overwhelmed with too many choices, an affect recognition-based tool (i.e. MRS) should be able to eliminate a lot of completely inappropriate pieces from consideration, helping the machine to whittle down the possibilities and present the user with a reasonable selection from which to choose. A typical MRS mainly has three components: a user, a music item, and an algorithm to match users with music items [39]. In this context, a common application for MER systems involves developing tools for end-users or groups of users.

Automatic categorization of music collections or particular pieces based on emotion and emotion-aware music recommendation has gained attention in the gaming and film industries, allowing them to generate a more immersive experience. Thus, an MRS could be presented as a software, application, or tool developed to suggest new songs or playlists to a user, basing its effectiveness on the degree of acceptance (likes/dislikes) expressed by the user and often using the user preferences as an input to generate the recommendations.

A comprehensive analysis of the evolution of music recommender systems over the past two decades is detailed in the study by Knees *et al.* (2020) [28]. The research defines distinct phases marked by significant advances that have shaped user interfaces. The study meticulously reviews multiple interfaces within each phase, elucidating their features, labels, dimensions, and display purposes. In one notable work highlighted by the study, Vad *et al.* (2015) [40] presents a probabilistic interface for music exploration and casual playlist generation. This interface leverages low-level audio descriptors to infer subjective features, employing a K-Nearest Neighbors (KNN) method for neighborhood selection in playlist generation. End-user evaluations gauge aspects such as perceived low-dimensional mood space projection, music exploration experience, and heat maps illustrating user interaction across the entire song universe. Schedl *et al.* (2018) [41] contribute to the discourse on challenges and visions in MRS research. Among the challenges discussed, the cold start problem arises from the lack of information and values when a user initiates interaction

with the software. Another challenge involves automatic playlist continuation, wherein additional tracks are added to a playlist while maintaining the target characteristics of the original playlist. The study highlights the difficulty of evaluating these systems, considering factors like user feedback and the elusive concept of serendipity, which pose challenges for direct integration in the system evaluation process.

2.2 Emotion Taxonomies

The concept of emotion is generally defined as a collection of psychological states that includes subjective experience, expressive behavior (i.e., facial, bodily, verbal), and peripheral physiological responses (e.g., heart rate, respiration) [42, 43]. Panda [44] states in its 2019 master's thesis that psychology researchers have discussed for a long time how emotions can be represented and classified due to their subjective nature. In the same work, emotions are described as diffuse reactions that vary from person to person, from moment to moment, and even across cultures. There are multiple words across different languages that are used to describe these emotional states, some of which are direct synonyms, while others represent slight variations and other elusive words that have no translation and are part of a specific culture.

In the work of Sing Tomar *et al.* [25], it is stated that the common aim of the emotional models is to describe, categorize, and model the "human emotional state". However, there is no commonly accepted emotion model; due to this, it is more likely that the researchers must choose and adapt one of the available models according to its application, research sub-field, and points of interest.

The relationship between music and emotion has been well-studied by psychologies for decades. The research problems faced by psychologists include whether the everyday emotions are the same as emotions that are perceived in music, whether music *represents* emotions (that are perceived by the listener) or *induces* emotions (the listener feels that), and how the musical, personal, and situational factors can affect emotion perception, directly concerning how music emotion should be conceptualized [45].

Among the MER literature, two predominant emotional taxonomy groups can be identified: *discrete* and *dimensional* [31, 46]. The importance of selecting the emotional taxonomy

to be used is that it defines the problem to approach. A discrete taxonomy implies the identification of emotional categories based on the selected features, which becomes a classification problem. A dimensional taxonomy must be addressed as a regression problem, as it involves predicting variable emotion ratings over time. However, an approach that is also taken by researchers in the area ([37, 47]) is to generate a regressor model that is then used as input for a classifier model that operates over the obtained results, allowing not only to have a finer granularity on the estimated values but also to obtain a general emotional category for the musical piece.

One of the problems with categorical approaches is the fact that there is no simple consensus about how many categories are needed to describe emotions in music. A problem that is shared with dimensional approaches, as the precision of dynamic MRS strongly relies on the number of the selected emotional labels and their distribution among the Valence/Arousal (V/A) plane [48], which is based in Russell's Circumplex model of Affect [49]. However, dimensional approaches also have a few limitations. Although the dimensional space permits the comparison of affect words to their reciprocal distance, it usually does not allow the formulation of operations between them. Furthermore, most emotional representations are not modeled on the fact that multiple emotions can be experienced concurrently [25].

2.2.1 Discrete

Kowalska & Wróbel in their work "Basic Emotions" [50], stated that the idea that there exists a small set of basic emotions dates back to the works of Descartes, who was first to suggest that all emotional states can be derived from six fundamental "passions": joy, sadness, love, desire, hatred, and wonder. However, the real debate supposedly begun with Darwin's book publication "The Expression of the Emotions in Man and Animal", where he argued that emotions are crucial for survival and thus they have distinctive expressions that all humans should accurately recognize. Based on this premise, the discrete or categorical approach to emotion conceptualization emerges by considering that people experience emotions as categories that are distinct from each other [45], relating them to the cognitive process that elicits emotion from predefined concepts or categories of emotions used

to model a person's emotional state [25]. These emotions are directly related to some emotional descriptors or adjectives that can be expressed as a finite set of "basic" or innate human emotions [51, 52] hard-wired in our brain and recognized universally [53], often associated with distinct patterns of physiological changes or emotional expressions [45].

Yang *et al.* [45] states that these categories, and particularly each "basic" emotion, can be defined functionally in terms of a key appraisal of goal-relevant events that have occurred frequently during evolution. These basic emotions can be found across many cultures, if not in all, and are often associated with distinct patterns of physiological changes or emotional expressions, a statement that can be related to Darwin's premise.

However, the major drawback of this approach is that the number of primary emotion classes is too small compared to the richness of music emotion perceived by humans. On the other hand, using a finer granularity does not necessarily solve this problem because the language for describing emotions is inherently ambiguous and varies from person to person. Moreover, using a large number of emotional classes or categories could overwhelm the subjects and is also impractical for psychological studies [45].

Eerola & Vouskoski [46] state that in studies that investigate music and emotions, Ekman's basic emotion model has often been modified to describe better the emotions that are commonly represented by music. For example, basic emotions rarely expressed by music, such as disgust, are often changed to more suitable emotion concepts like tenderness or peacefulness.

2.2.1.1 Ekman "basic emotions"

Presented in 1992 by Paul Ekman, and also currently known as the *basic emotion model*, it postulates that all emotions can be derived from a limited number of universal and innate basic emotions linked to distinctive universal facial expressions associated with them: anger, fear, sadness, happiness/joy, disgust, and surprise. It is built on the assumption that an independent neural system subserves every discrete basic emotion [54]. Later, in 1999, he further expanded his classification of emotions to include excitement, satisfaction, embarrassment, amusement, contempt, pride, and shame [55].

Ekman states that the adjective "basic" used to describe these emotions embodies two main

characteristics. One of these is that the emotions are discrete and can be distinguished fundamentally from one another. The second characteristic is the view that emotions have evolved through adaptation to our surroundings, derived from situations that have been useful in our ancestral environment. [56]. Then, these basic emotions, as suggested by Ekman, can be described by the following attributes [26]:

- Emotions are originated from innate instincts.
- Various individuals manifest the same emotion in response to the same circumstances.
- People tend to express basic emotions similarly.
- The physiological patterns of different people are consistent when experiencing basic emotions.

2.2.1.2 Hevner "Affective Ring"

The emotional model presented by Hevner in 1935 [57, 58], also known as the *Adjective List* or the *Affective Ring* model, is one of the earliest and most influential categorical music emotion models. In this investigation, Hevner conducted extensive experiments in which subjects were asked to report the adjectives that came to mind as the most representative part of a music piece was played. From these empirical studies, 67 emotional adjectives were found to describe the emotional space expressed by music. These 67 emotional adjectives can be divided into eight clusters or categories, namely dignified, sad, dreamy, serene, graceful, happy, exciting, and vigorous [35, 45]. The adjectives within each cluster are similar, whereas the meaning of the neighboring clusters varies cumulatively until a contrast in the opposite position is reached (Fig. 2.1).

2.2.2 Dimensional

Due to the limitation of the discrete models to express more subtle and rather complex affective states or emotional behaviors present in everyday interactions that people experience [59], the dimensional description of human affect has been gaining attention from the research community. In its 2011 book, "Music Emotion Recognition" [45], Yang describes dimensional emotion conceptualization as an area focused on the identification

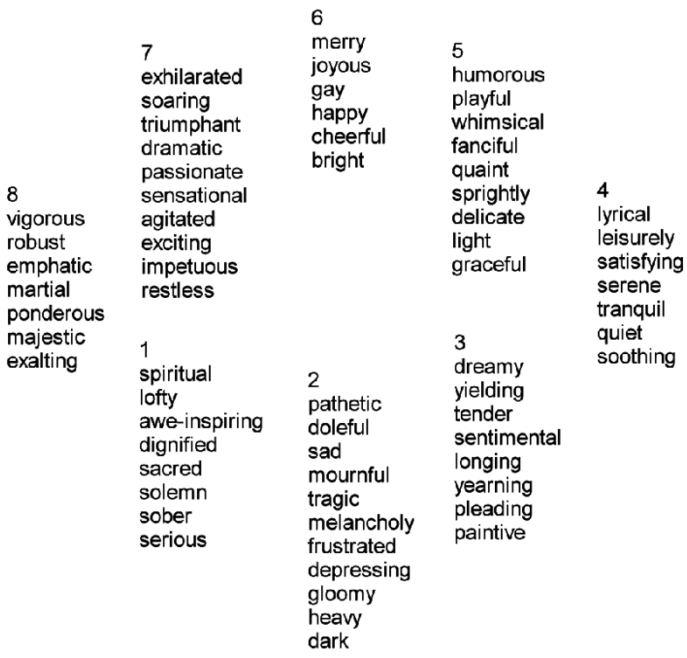


Figure 2.1: Hevner affective ring model [3].

of the emotions based on their positions on a small number of emotion dimensions with named axes, which are intended to correspond to internal human representations of emotion. This theory states that emotion should be depicted on a psychological dimensional space, and benefits by representing a wide range of emotions that not necessarily are bound to a specific emotion descriptor. Here, the similarity in the emotions can be expressed as a vector distance in the emotional space, and particularly, the two-dimensional models of emotion are based on the assumption that all emotional states arise from two independent neurophysiological systems, often referred to as Valence (or pleasantness; positive and negative affective states) and Arousal (or activation; energy and stimulation level). However, there is no consensus on the dimensions of emotion [52], as sometimes the Dominance (or potency; a sense of control or freedom to act) is considered too. It is also argued that the dimensional approach blurs important distinctions and consequently obscures important aspects of the emotion process depending on the approach. For example, anger and fear are emotions placed near the V/A plane on its second quadrant, but they have very different implications for the organism. The same applies to boredom and melancholy [45].

In their 2018 review work of features used in MER methods, Yang *et al.* [36] states that one of the main reasons that dimensional models are widely used in MER developments

is that this emotional model can be used to set up the corresponding relations between music features and emotional states simply and intuitively through the coordinates. Another statement drawn from their work points to the fact that Psychologists relate the arousal dimension to the *tempo* (fast or slow), *pitch* (high or low), *loudness level* (high or low), and *timbre* (bright or soft) music features, whilst the valence dimension is related to *mode* (major or minor) and *harmony* (consonant or dissonant).

2.2.2.1 Circumplex model of Affect

Presented in 1980 by Russell [49], this emotional model is based on the assumption that each basic emotion represents a bipolar entity being part of the same emotional continuum, where the emotions are considered as a mixture of two core dimensions: Valence and Arousal, orthogonally disposed in the *affective space* [51]. *Valence* is defined as the hedonic dimension of emotion, associated to a positive or negative degree related to a *pleasure-displeasure* continuum, ranging from pleasant to unpleasant. Arousal is defined as the mobilization of energy, which is associated with the intensity of the emotion or level of stimulation, related to *high/low* energy and ranging from calm to excited [60]. Figure 2.2 shows the emotional tags selected in this study over the V/A plane.

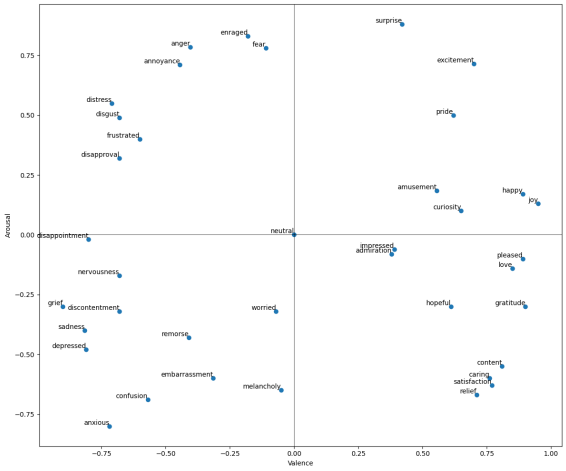


Figure 2.2: Two-dimensional circumplex space model. This emotional plot was adapted from a list of emotions and moods and their correspondent V/A values available in the works of Paltoglou & Thelwall [4] and Coutinho et al. [5].

2.2.2.2 Thayer's Mood model

Proposed by Robert E. Thayer in 1989 [61], this emotional model applies the circumplex model to music by implementing Energy and Stress measures. Here, Energy refers to the volume or intensity of sound in music, and Stress refers to the tonality and tempo of music. According to the stress and energy level, music mood can be divided into four clusters: Exuberance, Anxious, Contentment, and Depression (Fig. 2.3).

In this model, valence could be explained as varying combinations of energetic arousal and tense arousal. [46]

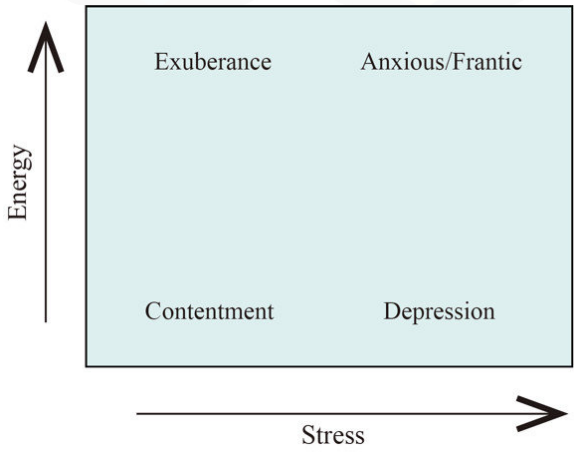


Figure 2.3: Thayer's model. Energy refers to the volume or intensity of sound in music, and Stress refers to the tonality and tempo of music. The mood is divided into four clusters: calm-energy (Exuberance), calm-tiredness (Contentment), tense-energy (Frantic or Anxious), and tense-tiredness (Depression) [6].

2.3 Emotional Trajectory

Due to the complexity and temporal variation of emotions in music, marking a piece with one annotation may be ambiguous and inaccurate. Therefore, dynamic emotion recognition must be done along the music to depict the flow of emotions expressed in music [62]. Music Emotion Variation Detection (MEVD), also known as dynamic MER [48], is defined by Yang *et al.* [37] as a research field that investigates the time-varying relationship between music and emotion, across different music styles. The objective of dynamic MER is to obtain emotion predictions for every short-time segment of a song, resulting in a series of emotion predictions and allowing users to track or visualize the emotion variation of a song as time unfolds [45]. However, there is a lack of a one-to-one or direct relationship between

music and emotional labels in the time domain. This is because the annotation values at a specific time point can be highly influenced by the music before that point, added to the bias induced by the annotators' psychological and physiological capabilities [62]. Thus, the emotion in music at a specific time can be expressed as the accumulation of a short piece of music before that point.

An approach to these estimations is by an *emotional trajectory* (emotion map, or simply a plot) of the valence-arousal values in time of a piece of music, which gives an insight into the emotions that are dominant in a composition, as well as their behavior over time, allowing to study both, emotional changes and tendencies [63]. It acts as a rough descriptor of the transitions in the emotional behavior of a musical piece, which can be used to find resemblances in the emotional trajectories of different songs, exploiting this applicability in MRS as the similarity criteria. The development of a model capable of predicting an approximate emotional trajectory of a song, full or excerpt, by identifying the transitions among emotions could lead to the implementation of an Emotional Music Recommendation System based on querying a particular emotion dynamic over time, generating then a playlist of songs that share similar approximate emotional distributions.

Weniger *et al.* [64] states that continuous dimensional emotion recognition from audio is inherently a regression problem. Due to this, it aims to maximize the correlation between sequences of regression outputs and the continuous-values emotion contours while minimizing the average deviation between them. Therefore, the ultimate goal of an emotional trajectory estimator is to predict and track the changes of emotion within music over time [38]. In Warmbrodt *et al.* work [65], the concept of *emotional trajectory* is presented as the *order of emotional music*, related to determining *how successfully music evokes desired emotions* or *how successfully the musical mood can be perceived by the listener* depending on the approach.

2.4 Datasets

2.4.1 DEAM

The MediaEval Database for Emotional Analysis of Music (DEAM) [66] is one of the largest publicly available datasets in continuous MEVD. It was developed within the Emotion in Music (EiM) Task of the Multimedia Evaluation Campaign (MediaEval) between 2013 and 2015 [7, 67–70], as a public benchmark and evaluation framework to standardize the comparison between MER methods. It consists of 1,802 royalty-free audio files (58 full-length songs and 1,744 excerpts of 45 seconds) that cover a variety of Western popular music genres: *rock, pop, soul, blues, electronic, classical, hip-hop, international, experimental, jazz, country, rap, reggae,* and *world*.

The static and dynamic arousal and valence annotations available for each song were obtained by two different approaches, in a laboratory and through a crowd-sourcing approach using Amazon Mechanical Turk (MTurk). For the second approach, participants had to use a web interface (Fig. 2.4) to register valence and arousal ratings for each song after it had been listened to at least once. These annotations were registered using a scale from -10 to 10 for the dynamic approach, and a nine-point Self-Assessment Manikin [71] scale for the overall (i.e., static) values. The annotations consistency was measured by applying Cronbach's α test on the sequences of annotations of each song and computing the coefficient of determination of a Generative Additive Model that generalizes the song's annotations across annotators [7]. For this, the annotations were resampled to 2 Hz and then normalized by adding the delta obtained between the mean of the annotations by an annotator and the global song annotation mean for that song timestamp. The dataset also has available metadata about the songs and the annotators, such as the genre labels of the songs, the familiarity, and the liking of each of the annotated songs for each annotator, and even the personality traits of the annotators (for 2014 values). Although the dataset is mainly designed for dynamic MER, the annotations can also be useful for static MER after being summarized over time.

The 2013 MediaEval edition [67, 72] included the *brave new task* of Emotion in

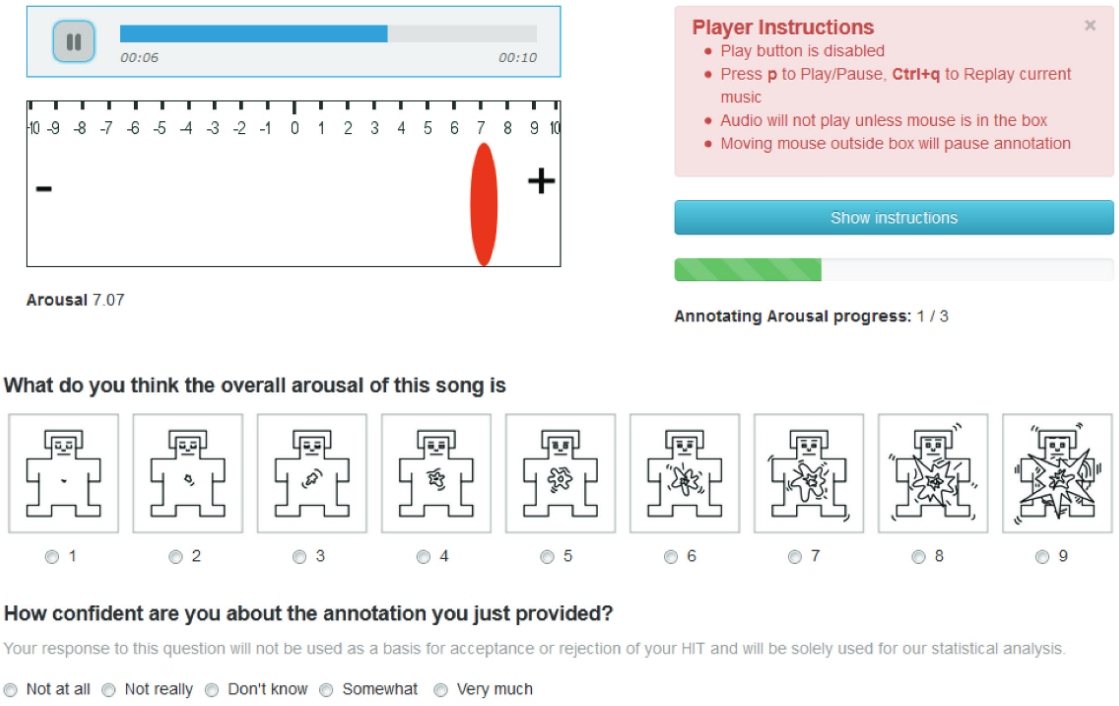
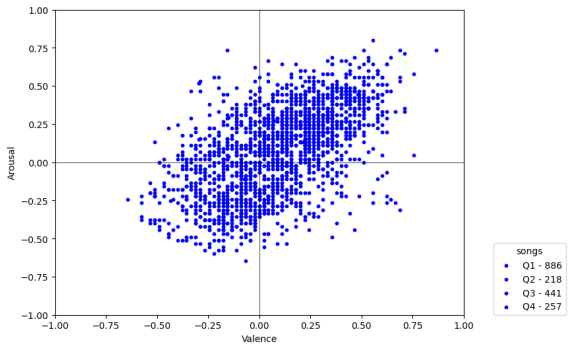


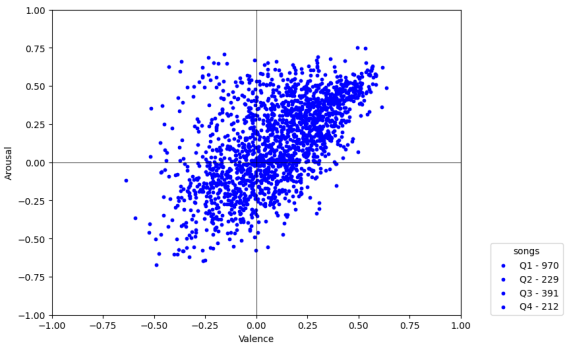
Figure 2.4: Web annotation interface used for registering dynamic and static arousal and valence ratings for each song after it has been listened to at least once [7].

Music, which consisted of two sub-tasks: *dynamic* and *static* emotion characterization. The training dataset consisted of 700 excerpts of 45 seconds, labeled with static and dynamic annotations, where the dynamic ones were obtained in a 1Hz time resolution scale. The evaluation set consisted of the 300 remaining clips. This dataset version is also known as the 1,000 *songs dataset* of the EmoMusic Database [73] and is commonly referred to as such by researchers in the area. Then, in the 2014 edition [68], the static sub-task was replaced by a feature design sub-task that encouraged the participants to develop new features, with a limitation of adding only one feature per experiment to the baseline feature set, and only 5 runs that they can submit. However, this task was not very popular, as only one team submitted results [74]. Finally, in the 2015 workshop edition [69], a subset of 431 songs was selected from among 1,744 songs to generate the training set for the challenge; meanwhile the temporal resolution for the values of the dynamic task was changed to 2Hz. The evaluation set consisted of 58 complete music pieces, with an average duration of 234 ± 105.7 seconds. The feature set is composed of 260 low-level features obtained with the *openSMILE toolbox* [1] from non-overlapping segments of 500ms, with a frame size of

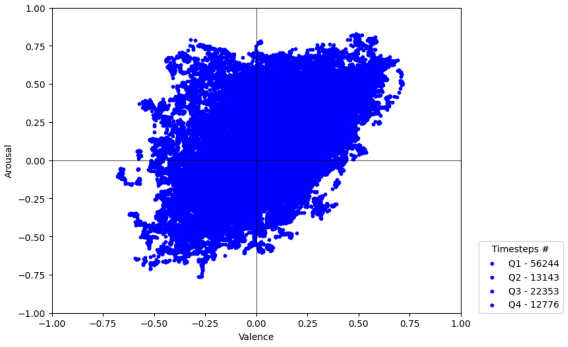
60ms and 10 ms step. These features comprise 65 low-level descriptors (LLDs) and their first-order derivatives (130 LLDs total). The mean and standard deviation of each LLD over 1s time windows with 50% overlap (0.5s step size) are also calculated to adapt the LLDs to the challenge requirements, resulting in 260 low-level features exported at a 2Hz rate (Table 2.1). Figure 2.5 displays the available annotations data for a static or dynamic approach and a third case for an averaged dynamic approach, similar to the static case but obtained by averaging the time annotations for each song.



(a) Static df



(b) Dynamic avg df



(c) Dynamic set df

Figure 2.5: DEAM V/A values distribution per song for each approach: static, dynamic avg., and dynamic.

Metrics

The evaluation process of the models implemented over the DEAM dataset compares the predicted values (Y) with the ground truth annotation values (X). This comparison is performed at a song level for the dynamic approach and then averaged across songs. Aljanaki *et al.* [7] selected the following metrics to evaluate the models in the dataset:

- Pearson's correlation coefficient (PCC) ρ: It measures the linear relationship between

the predicted and reference values, expressed by:

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{2.1}$$

where $Cov(X, Y)$ is the covariance of X and Y and σ_X is the standard deviation of X.

- Root Mean Square Error (RMSE): It represents a standard deviation of the differences of the predicted values from the line of best fit at the sample level. Given N predicted samples Y_n the corresponding reference samples X_n, then the RMSE between them can be written as follows:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N} (Y_n - X_n)^2}{N}} \tag{2.2}$$

- Concordance Correlation Coefficient (CCC) [75] ρ_c, defined as:

$$\rho_c = \frac{2Cov(X, Y)}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2} \tag{2.3}$$

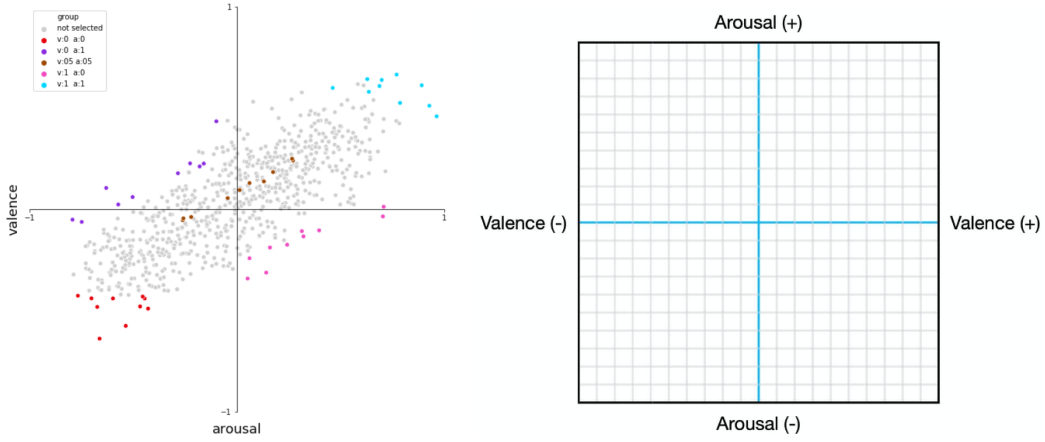
where X and Y are the vectors of numbers to compare (ground truth and predictions, respectively), σ_X^2 is the variance of X, $Cov(X, Y)$ is the covariance of X and Y, and μ_X is the mean of the vector X.

These measures offer insights into the prediction capabilities of models by assessing similarities and correlation behavior. In this context, RMSE measures how accurately the predicted emotion annotation values align with the true song values, while ρ indicates the correctness of the guessed change direction. Conversely, the CCC metric evaluates agreement between these two time series by scaling their correlation coefficient with their mean square difference, assessing the trend shape of the annotation traces (emotional trajectory). hence, annotation predictions that correlate well with the gold standard but shift in value are penalized in proportion to the deviation. This metric has been stated as the metric of choice for continuous emotion recognition and has been utilized in other continuous emotion recognition tasks, such as the Audio/Visual Emotional Challenge and Workshop (AVEC) since its 2015 edition [76–78].

2.4.2 MERP

The Music Emotion Recognition with Profile information (MERP) dataset is a publicly available dataset that combines dynamic affect labels of full-length musical pieces with participant profile information [8]. The primary goal is to investigate whether adding listener profile information can enhance MER. The dataset comprises 54 full songs, with 50 selected from the *top* $1,000$*songs listened to* list of the Free Music Archive and filtered to ensure a representative amount of data (Fig. 2.6a). The remaining four songs were selected from the DEAM dataset.

A dynamic approach was used to obtain MERP arousal and valence labels through MTurk crowdsourcing. Participants simultaneously labeled VA values while listening to stimuli, with mouse movement captured over a two-dimensional VA graph at a frequency of 10 Hz (Fig. 2.6b). Dynamic labels were compared to DEAM dynamic annotations to evaluate annotation quality and filter participant entries. The collected profile information was categorized into three sections: demographic information, listening preferences, and musical experience. Demographic information included age, gender, country of residence, and country of musical enculturation. Listening preferences included preferred languages and favorite music genres, while musical experience covered instrument playing and formal music training.



(a) **(b)**

Figure 2.6: (a) Representation of the predicted arousal and valence value of the top 1000 songs of the FMA (filtered by length criteria). Colored dots represent the songs selected for their investigation. (b) Annotation interface on which arousal and valence values were captured via mouse tracking in the listening study [8].

For benchmarking, the authors provided baseline prediction models for arousal and valence. Two model architectures were explored: fully connected and LSTM models (Fig. 2.7). The models used audio features extracted by openSMILE toolbox and profile features binned into categories. Separate models were built for each profile feature to identify their impact on prediction accuracy.

Two variations of the models were trained: one using audio features alone and one using audio features concatenated with a single profile type feature. The models were trained using 5-fold cross-validation, with inputs set to 30 timesteps and trained for 100 epochs. The reported configuration parameters included the Adam optimizer with a learning rate of 0.0001 and a batch size of 8. Metrics included Mean Squared Error (MSE) and Pearson Correlation Coefficient (ρ) between predicted and ground truth V/A values. The dataset [79] as well as the model code [80] are available online.

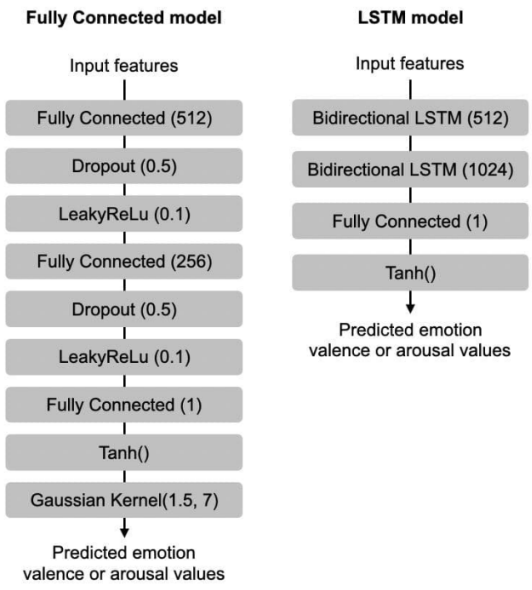


Figure 2.7: Presented baseline architectures [8]. **Left**: Fully Connected architecture. **Right**: LSTM architecture.

2.4.3 MUSAV

The MusAV dataset, introduced by Bogdanov *et al.* in 2022 [9], serves as a public benchmark for the comparative validation of arousal and valence regression models in audio-based music emotion recognition. This dataset is constructed through comparative annotations of arousal and valence on pairs of tracks (Fig. 2.8), utilizing audio previews

and metadata retrieved from the Spotify API. The MusAV dataset encompasses 2,092 track previews spanning 1,404 genres. The models presented to validate their proposal consisted of Convolutional Neural Network (CNN), Visual Geometry Group (VGG), and EfficientNet-based approaches. These models were initially trained on the DEAM, EmoMusic, and MuSe datasets in different experiments and subsequently tested over the newly introduced MusAV dataset. The authors reported respective RMSE scores for each dimension and each trained model.

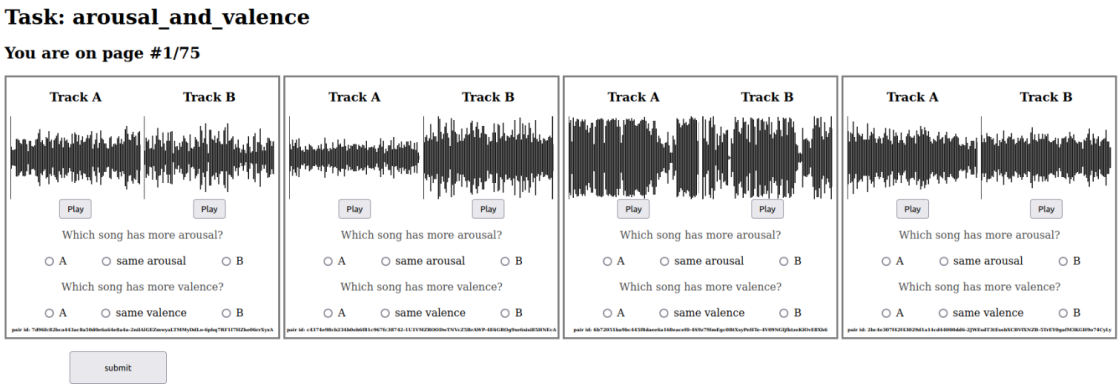


Figure 2.8: Web annotation interface employed for recording comparative arousal and valence ratings for each song. Figure adapted from [9].

2.4.4 AVEC

The Audio/Visual Emotion Challenge and Workshop (AVEC) is established with the primary objective of providing a standardized benchmark test set for multimodal information processing. This initiative aims to unite the depression and emotion recognition communities. The central motivation behind AVEC is to propel emotion recognition and depression estimation for multimedia retrieval to a stage where behaviors expressed in human-human or human-agent interactions can be accurately and consistently detected in real-life conditions. Key references to AVEC include Ringeval *et al.* works in 2015 [76] and 2017 [78], and Valstar *et al.* 2016 work [77], which encompasses the version of the benchmark reviewed for this investigation. The 2017 version of the challenge was centered around the Distress Analysis Interview Corpus of human and computer interviews (DAIC-WOZ) data set [81] and the SEWA [82] dataset. The SEWA dataset captures spontaneous and naturalistic human-human interactions in the wild, offering audio and visual modalities for

a comprehensive evaluation of emotion and distress recognition algorithms.

For the Affect Sub Challenge (ASC) within AVEC, participants are tasked with performing fully continuous affect recognition of three affective dimensions: Arousal, Valence, and Likability. The goal is to predict the level of affect for every moment in the recording. The competition measure employed is the *concordance correlation coefficient (CCC)*, and this sub-challenge utilizes the data available in the SEWA dataset.

2.5 Feature Engineering

In general terms, and as defined by Panda *et al.* (2023) [83], a feature is a characteristic part of something. Features help to distinguish one thing from another, by providing the essential descriptive primitives by which individual objects or works may be identified. In musical terms, features may be characteristic of a musical work, of a movement, of a composer, of a very specific musical dimension, of a genre, and so forth. Then, the goal of feature extraction is to reduce the information of songs to descriptors that can accurately describe them. Emotionally relevant audio features for MER can be grouped into three categories: low-level (e.g., spectral features, MFCC, etc.), perceptual (e.g., rhythm, clarity, modality, articulation, etc.), and high-level semantic features (e.g., genre, danceability, etc.).

Among the developed tools for the extraction of musical features, the Munich open-source Media Interpretation by Large feature-space Extraction (openSMILE) was presented in 2010 as a novel open-source feature extractor toolkit designed for incremental processing, uniting the feature extraction algorithms from the speech processing and the Music Information Retrieval Communities [1]. Although its primarily intended area of use is audio feature extraction, it is stated to be principally modality independent, e.g. physiological features such as heart rate, EEG, or EMG signals can also be analyzed with openSMILE using audio processing algorithms. The documentation of openSMILE [84] describes the main Low-Level Descriptors (LLD) that this toolkit is capable of extracting (Table 2.1). The authors provide extra details about two features whose extraction process can be configured depending on the desired information details:

- Chroma features: 12 semitones are computed from a short-time FFT spectrogram (window-size 50[ms], rate 10[ms], Gauss-window). The spectrogram is scaled to a semi-tone frequency axis using triangular filters. The extractor also provides examples of how to obtain a single vector that contains the mean value of the Chroma features computed over the complete input sequence. Such a vector can be used to recognize the *musical key* of the song.

- MFCC features: Mel-Frequency Cepstral Coefficients from 25[ms] audio frames (sampled at a rate of 10[ms]) (Hamming window). 13 (0-12) or 12 (1-12) MFCCs are computed from 26 Mel-frequency bands, and a cepstral liftering filter with a weight parameter of 22 is applied. 13 delta and 13 acceleration coefficients are appended to the MFCC. Then, two other configurations are presented for 0/1-12 MFCC computation, where the features are now mean normalized with respect to the full input sequence. The frequency range of Mel-spectrum is set from 0 to 8[kHz].

Table 2.1: OpenSMILE's low-level descriptors [1].

Feature Group	Description
Waveform	Zero-Crossings, Extremes, DC
Signal energy	Root Mean Square & logarithmic
Loudness	Intensity & approx. loudness
FFT Spectrum	Phase, magnitude (lin, dB, dBA)
ACF, Cepstrum	Autocorrelation and Cepstrum
Mel/Bark spectr.	Bands $0 - N_{mel}$
Semitone spectr.	FFT based and filter based
Cepstral	Cepstral features, e.g. MFCC, PLP-CC
Pitch	F_0 via ACF and SHS methods, Probability of Voicing
Voice Quality	HNR, Jitter, Shimmer
LPC	LPC coeff., reflect. coeff., residual, Line spectral pairs (LSP)
Auditory	Auditory spectra and PLP coeff.
Formants	Center frequencies and bandwidths
Spectral	Energy in N user-defined bands, multiple roll-off points, centroid, entropy, flux, and rel.pos. of max./min.
Tonal	CHROMA, CENS, CHROMA-based features

3 | Literature review

This chapter offers a comprehensive technical overview of the deep learning models reviewed in the investigation, giving an insight into the properties and capabilities exploited by researchers in the developed models. It explores diverse implemented techniques for MER applications, detailing the selected model approach, emotional taxonomy, and datasets employed in the research endeavors. Through this exploration, the methodologies and frameworks reviewed in this investigation are elucidated, providing valuable insights into the landscape of MER research. The reviewed techniques encompass classical regression methods, which laid the groundwork for early approaches in the research field; convolutional models spanning from simple to intricate architectures aimed at enhancing feature extraction; recurrent networks that exploit the temporal nature of dynamic music emotion recognition challenges; and Transformer-based models that aim to further expand the modalities and to take advantage of the model capabilities. Then, the selected benchmark models are presented, accompanied by brief explanations of their behavior.

3.1 Prediction Approaches for MER systems

3.1.1 Classical Techniques

The literature review process was primarily focused on models that surpass traditional RNNs in terms of implementation and characteristics. Nonetheless, among the reviewed techniques, some more classical approaches offer intriguing perspectives relevant to the research focus:

- Kim *et al.* (2011) [85]: This work presents a music mood classification model based on valence-arousal values for music recommendation. Utilizing a K-means clustering algorithm and a self-made dataset, 8 emotional clusters comprising multiple emotional tags in the V/A plane were obtained and compared with Hevner's emotional categories to validate the clustering approach.

- Medina *et al.* (2019, 2022) [13, 86]: In these works, an automatic MER prediction system based on a Multilayer Perceptron (MLP) Regressor. The study covers the design process and the evaluation of model configuration parameters through various experiments.

- Krols *et al.* (2023) [87]: This study presents a Multiple Linear Regression method for uni and multimodal prediction of VA scores from the Deezer Mood Detection Dataset. It utilizes available high-level Spotify API features along with mood annotations obtained from LastFM. The model's performance was compared with Support Vector Regression (SVR), MLP, and Random Forest Regressor to validate the set of features and the behavior of the fused modalities.

3.1.2 Convolutional Models

Following the previous approach, an overview of various convolutional-based models applied in emotion recognition tasks is presented:

- Sarkar *et al.* (2020) [47]: Developed a CNN model that receives spectrogram representations of the data. The proposed architecture is built around VGGNet, and the Soundtrack dataset is used to validate the results by comparing the main four categories corresponding to Russell's four quadrants: *happy*, *anger*, *sad*, and *neutral*.

- Allognon *et al.* (2020) [88]: Implemented a Deep Convolutional Autoencoder method based on CNN to train an SVR approach for facial expression recognition. The experimental results were reported by testing the method over the RECOLA 2016 dataset, with the CCC metric as the evaluation metric.

- Malik *et al.* (2017) [89]: Presented a CNN-RNN architecture obtained by stacking a

CNN with two FC branches and one RNN for each V/A dimension. This model maps the input feature vectors into valence and arousal values. Another model variation was presented without branching, thus training both dimensions on the same branch.

- Dong *et al.* (2019) [90]: Introduced a Bidirectional Convolutional Recurrent Sparse Network (BCRSN) that extracts and learns features from the spectrogram of the input songs. The performance evaluation is conducted on the DEAM dataset, with results presented for valence and arousal dimensions including CCC metrics.

- Liu *et al.* (2017) [91]: Presented a CNN-based method that uses spectrograms as input, as means to avoid "complex artificially selected features". Experiments are conducted on the CAL500 dataset.

- Orjesek *et al.* (2019, 2022) [48, 92]: Proposed a recurrent RNN architecture for the recognition of music directly from the raw Pulse-code Modulated audio signal. The presented architecture consists of two main parts: a short-term audio feature extractor and an RNN that captures the temporal variations of the features. The feature extractor feeds a 1dCNN layer stacked with an autoencoder-based iterative reconstruction (IR) unit. Authors indicate that Back-Propagation Through Time (BPTT) and Adam optimizer were selected for model training. Then, to evaluate the resulting model, DEAM dataset was used. Metrics PCC and RMSE are reported, showing that an end-to-end approach outperforms hand-crafted feature approaches and indicating that commonly used audio features describe the valence dimension inefficiently.

3.1.3 Recurrent Networks

Long Short-Term Memory (LSTM) is a highly functional sequential sequential model that represents a redesign of Recurrent Neural Networks (RNN). By incorporating input, forget, and output gates within the memory block, LSTM excels in exploiting and retaining information over longer time spans compared to traditional RNNs. As a result, researchers have suggested that since LSTM models are proficient at capturing and storing information for extended durations, bidirectional-LSTM (BLSTM) approaches can effectively access context in both preceding and subsequent directions, thus enhancing its predictive

capabilities [93]. Several approaches were reviewed, including:

- Li *et al.* (2016) [93]: Proposes a Deep Bidirectional LSTM (DBLSTM) based multi-scale regression method tasked with Dynamic MER. The model aims to map the feature sequence to a sequence of V/A values. Inspired by deep-feed networks, the authors introduced a stacked multi-scale recurrent hidden layers approach to enhance data representation, resulting in the DBLSTM model. Reported results demonstrate competitive performance compared to state-of-the-art methods.

- Ma *et al.* (2017) [62]: Introduces a Multi-scale Context Attention (MCA) fusion method, building on Li *et al.* multi-scale approach. Extending the attention mechanism to a multi-scale context fusion, the model employs LSTM and Attention-based LSTM (A-LSTM) layers to obtain context vectors. V/A value prediction is based on the weighted sum of multi-scale context vectors. The best results were achieved with the A-LSTM MCA approach, validating the effectiveness of Attention-based models. However, this model only considers preceding sequence content, potentially overlooking valuable subsequent information about music emotion.

- Liu *et al.* (2019) [94]: Presents a Bi-RNN (LSTM) model to extract features from music chroma spectrum. Experiments conducted on the DEAM dataset utilize a categorical approach, assuming excited-calm and joy-sad emotional dimensions.

- Sun *et al.* (2020) [95]: Introduces a multimodal LSTM-based model with a self-attention mechanism to enhance LSTM's ability to capture longer temporal dependencies. Experiments on the MuSe 2020 dataset report unimodal and multimodal results using MSE and CCC metrics.

- Zhang *et al.* (2023) [96]: Presents a Dual Attention-based Multi-scale Feature Fusion (DMAFF) method alongside the development of the MER1101 dataset used in their work. Utilizing the DAFF module, the model achieves multi-scale context fusion and captures emotion-critical features in spatial and channel dimensions. Evaluation metrics including CCC, PCC and RMSE compare model capabilities over the DEAM dataset.

3.1.4 Transformer

3.1.4.1 Model Overview

Introduced by Vaswani *et al.* [10] for machine translation tasks, the Transformer technique is built upon the *Attention Mechanism*, enabling parallelization and establishing a global relationship between input and output [97]. Unlike traditional attention models, the Transformer employs Multi-Head Attention Mechanisms, which enhance its performance across various Natural Language Processing (NLP) tasks by interpreting word sequences. This capability, combined with the model's proficiency in capturing long-range dependencies and interactions, makes it particularly appealing for time series modeling. Consequently, Transformer models can function as auto-regressive models, akin to LSTMs, enabling their use in generative tasks where attention mechanisms play a crucial role in building contextualized signal representations, prioritizing relevant segments [98].

Several Transformer-based approaches have been developed for unimodal and multimodal emotion recognition, taking advantage of the Self-Attention mechanism to relate different positions within a sequence and compute representations. While in classic convolutional neural networks, the number of operations required to relate signals between arbitrary input or output positions increases with the distance between them, the Transformer model architecture, along with the Multi-Head Attention (MHA) mechanism, mitigates this by reducing the operations to a constant number. However, this may result in reduced effective resolution due to averaging attention-weighted positions, a limitation addressed by the MHA mechanism. Figure 3.1 displays the architecture of a typical Transformer models presented by Vaswani *et al.*

Attention

An attention function can be described as mapping a query and a set of key-value pairs to an output where the query, keys, values, and output are all vectors. The output is considered a weighted sum of the values, where a compatibility function of the query with the corresponding key computes the weight assigned to each value.

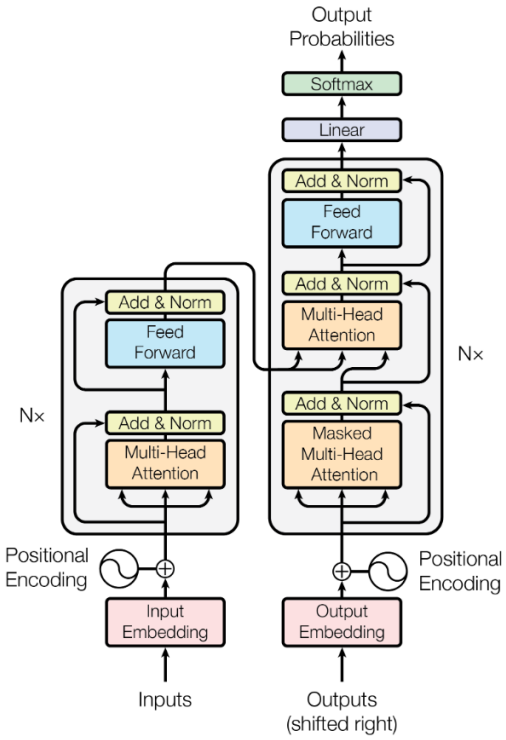


Figure 3.1: Transformer model architecture presented by Vaswani *et al.* [10].

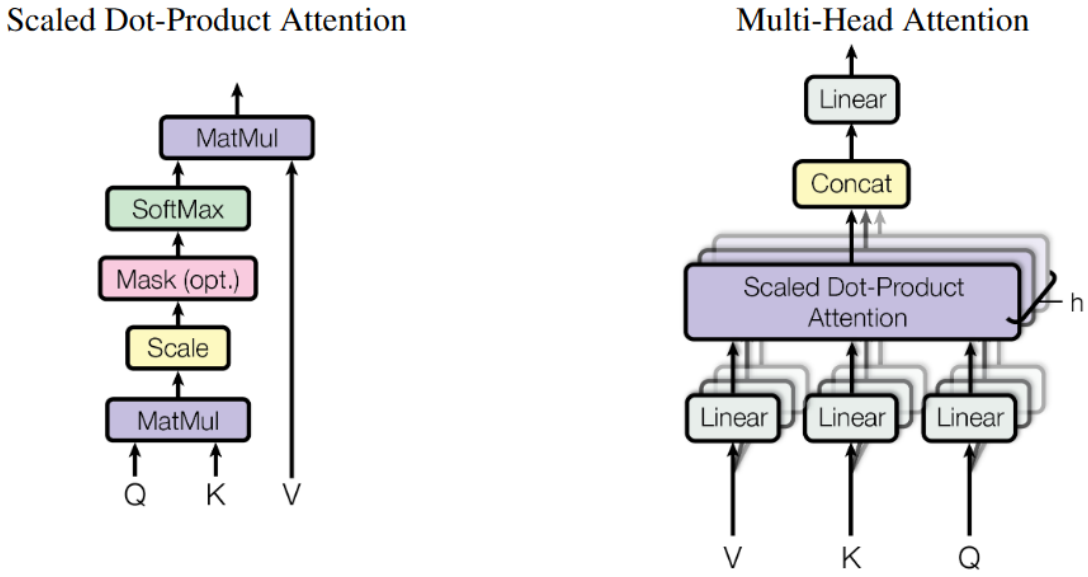


Figure 3.2: (left) Scaled Dot-Product Attention. (right) Multi-head attention consists of several attention layers running in parallel. Vaswani *et al.* [10].

Scaled Dot-Product Attention

Scaled Dot-Product Attention has three inputs: queries and keys of dimension d_k, and values of dimension d_v. A dot product is computed over the query with all keys, divided by $\sqrt{d_k}$, and a softmax function is then applied to obtain the weights on the values. In practice, the attention function is computed simultaneously on a set of queries packed together into a matrix Q. The keys and values are also packed into matrices K and V. Then, the matrix of outputs is computed as:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k})V \tag{3.1}$$

The scaling factor $\frac{1}{\sqrt{d_k}}$ was presented as a means to counteract the possible extremely small gradient values that the softmax function could reach for large values of d_k.

Multi-Head Attention

Instead of performing a single attention function with d_{model}-dimensional keys, values and queries, the authors perform a linear projection of the queries, keys and values h times with different learned linear projections to d_k, d_k and d_v dimensions respectively. Then, on each of these projected versions, the attention function is performed in parallel, yielding d_v-dimensional output values. These are concatenated and once again projected, resulting in the final values (Fig. 3.2). MHA allows the model to jointly attend to information from different representation subspaces at different positions, which can be expressed as:

$$\begin{aligned} MultiHeadAttn(Q, K, V) &= Concat(head_1, \ldots, head_H)W^O \\ \text{where } head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V). \end{aligned} \tag{3.2}$$

The projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$.

3.1.4.2 Highlighted Works Reviewed

This literature review provides comprehensive insights into various transformer-based models applied to both unimodal and multimodal emotion recognition tasks. Each model is

designed to address different modalities and tasks, showcasing the versatility of transformer architectures across diverse facets of emotion recognition. The review encompasses a spectrum of works, ranging from those focused on preprocessing ECG signals for continuous emotion recognition to hybrid LSTM-Transformer architectures for Speech Emotion Recognition (SER) tasks. The multimodal context includes models integrating audio, visual, text, and physiological features for comprehensive emotion prediction.

Unimodal Transformer Approaches

- Vazquez-Rodriguez *et al.* (2022) [98]: Explores a transformer encoder-based model for preprocessing ECG signals, primarily for a binary emotion recognition task, predicting high/low levels of arousal and valence. Contextualized representations from ECG signals obtained from the AMIGOS dataset ar the primary focus.

- Andayani *et al.* (2022) [97]: Proposes a hybrid LSTM-Transformer architecture for SER, replacing positional encoding with an LSTM layer. Operating on Mel Frequency Cepstral Coefficient (MFCC) features, the model aims for improved emotion prediction.

- Agrawal *et al.* (2021) [99]: Presents a lyric-based categorical MER approach. Using databases designed over the Russell V/A model, the model predicts the emotional quadrant of songs based on English lyrics. The architecture, based on XLNet, employs a large bidirectional transformer with improved training methodology, an extension of BERT improved by a Transformer XL architecture.

Multimodal Transformer Approaches

- He *et al.* (2022) [100]: Introduces a Transformer-based multimodal temporal attention model (MMTA), processing features from audio, visual, text, and physiological modalities. Features are extracted, projected onto a linear layer, passed through a Temporal Convolutional Network (TCN) capturing local temporal context, and fused using a transformer for multimodal fusion and long-range temporal context aggregation. Experiments were performed as part of the MuSe-stress sub-challenge

in the MuSe 2022 Multimodal Sentiment Analysis Challenge [101]. The unimodal performance of the visual and audio feature sets achieved sufficient performance in arousal and valence dimensions to extend it to multimodal performance analysis to ensure prediction performance and model robustness.

- Zhao *et al.* (2023) [102]: Presents a transformer-based deep-scale fusion network (TDFNet) for multimodal emotional recognition, emphasizing deep-scale features and speaker-related mutual correlations between audio and text to predict emotion values. Evaluation on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset predicts four emotion categories *sad, happy, angry, and neutral*. Results indicate that the transformer-based model can generate a discriminate representation of emotions in multimodal emotion recognition.

- Guo *et al.* (2022) [103]: Unveils an Implicitly Aligned Multimodal Transformer Fusion (IA-MMTF) model implementing bidirectional LSTM layers for audio and text representations. It employs cross-modal attention to learn implicit alignments between modalities and uses a weighted fusion layer to obtain complementary emotional representations. Evaluation over IEMOCAP dataset validates the model's unimodal and multimodal setup approaches.

- Hsu & Wu (2023) [104]: Introduces a bimodal Transformer Encoder with Segment-Level Attention for audio-visual emotion recognition, rooted in Ekman's categorical model of emotions. Their approach presents a Neural Tensor Network to calculate emotional consistency scores of audio and visual features in the same segment.

- Sun *et al.* (2023) [105]: Proposes a tensor-based multimodal Transformer framework with a global attention mechanism for depression detection. It incorporates textual, acoustic, and visual features, utilizing CCC and RMSE as comparison metrics.

- Huang *et al.* (2020) [14]: Advocates a transformer-based multimodal model incorporating an architecture without and encoder-decoder structure. It employs a linear layer for audio and visual inputs transformation, achieving effective performance in dynamic emotion recognition, with additional exploration of LSTM layers. This

model trained on the audio and visual modalities of the AVEC multimodal database [78] demonstrate effective performance, particularly for the arousal dimension.

3.2 Prediction Approaches of Emotional Trajectory

This section aimed to review various MER models and techniques based on Russell's V/A models that presented, used, or resembled the concept of an *emotional trajectory*. However, among the reviewed works, only one research presented concepts similar to those outlined in this work.

In Grekow's works [63, 106], the concept of an emotion map is presented as a means to describe the emotional behavior of a piece of music. This research is grounded in a self-generated dataset comprising 324 six-second fragments of diverse musical genres, each annotated with V/A values by five music experts with university-level musical education. The primary focus was to experiment with different features (low-level, rhythm, and tonal) and combinations thereof to enhance predictions using regressors.

3.3 Benchmark

3.3.1 DEAM

The working notes submitted by the teams that participated in the 2015 edition of the MediaEval workshop are available on its website [2]. These works were reviewed to compare their submissions' techniques, approaches, and results. All of their models used the development set that contained 431 excerpts of 45 seconds each, the 58 full songs as an evaluation set, and the 260 low-level spectral features obtained with openSMILE. Reported results of these models are resumed in Table 3.1.

Aljanaki *et al.* [107] presented a comparative approach between the baseline feature set and their own feature set obtained with Essentia open-source framework [112] which contains 40 low-level and high-level features (i.e. scale, tempo, tonal stability), and uses bigger time windows for feature extraction. A Gaussian Process (GP) regression is applied to predict the valence and arousal values per segment, using a maximum likelihood estima-

Ref.	Method	Reported RMSE	
		Arousal	Valence
[69]	Multiple Linear Regression (MLR)	0.270±0.11	0.366±0.18
[107]	Gaussian Process	0.285±0.124	0.295±0.147
[108]	LR+S	0.24	0.35
	LSB+S	0.24	0.35
[109]	RNN	0.247±0.116	0.365±0.188
[110]	LSTM-RNN	0.242±0.116	0.373±0.195
[111]	BLSTM-RNN	0.230±0.11	0.331±0.18
	BLSTM-ELM	0.234±0.11	0.308±0.17
	SVR	0.250±0.15	0.303±0.19

Table 3.1: RMSE metrics for Arousal and Valence dimensions. Models available in the working notes papers of MediaEval 2015 Emotion in Music Task [2].

tion of the best set of parameters. Their proposed feature set showed better performance for arousal in terms of correlation and RMSE, but worse for valence. Both obtained models were reported to perform *unacceptably bad* on the valence dimension.

In Gupta *et al.* [108] work, three different regression methods based on the 260 feature baseline set are presented. The first model consisted of a Linear Regression with a Smoothing step (LR+S) performed by low-pass filtering the frame-wise arousal and valence values obtained for each dimension. The smoothing operation was stated to not only remove high-frequency noise but also to incorporate local context, since the decision for a frame is given by an unweighted combination of frame values in a window centered around that frame, thus incorporating local context. The next model was a Least Squares Boosting trained with a Smoothing step (LSB+S) for each dimension, with the same smoothing operation as before. Finally, an unweighted combination of the previous models with a model based on a Boosted Ensemble of Single Feature Filters (BESiF) was also reported. Their approach fails for valence prediction, obtaining close to no correlation with the ground truth values

A Recurrent Neural Network (RNN) was implemented by Pellegrini *et al.* [109] with fine-tuned parameters obtained with a 10-fold cross-validation experiment. For their experiments, they used the baseline feature and a custom feature set that included hand-picked features obtained with the Essentia framework to train different RNN models, which were also tested in two different setups: the application of a moving average filter to smooth the predictions and the addition of a second RNN layer that used the output of the first layer as input. However, as the task was limited in the amount of trials to submit by the team,

the results of the second RNN with smoothing over the 260 baseline feature set were not reported. Their RNN approach with smoothing was shown to be slightly better than the baseline values.

A Long-Short Term Memory Recurrent Neural Network (LSTM-RNN) approach was presented in Coutinho *et al*. work [110]. An LSTM-RNN network is similar to an RNN except that the nonlinear hidden units are replaced by a special kind of memory blocks which overcome the vanishing gradient problem of RNNs. A multi-task learning framework was used for joint learning of arousal and valence time-continuous values. The architecture presented consisted of an LSTM-RNN with 3 hidden layers, where the first layer was pre-trained under a de-noising auto-encoder (DAE). The development and test sets presented in the 2014 edition of MediaEval were used to train the DAE. The results presented consisted of the mean of a number of LSTM-RNN outputs selected from a pool of 10-fold cross-validation trials. Their results showed slightly better results than the baseline values.

The approach taken by Xu *et al.*[111] was to evaluate and implement several multi-scale methods at three different levels, including the acoustic feature level, the regression model level, and the emotion annotation level. The feature set was organized into groups according to their time scales and fundamentals for the acoustic feature level, resulting in 4 feature sets with different dimensions. A bi-directional LSTM-RNN (BLSTM-RNN) method was implemented and trained with different time scales (60, 30, 20, and 10) on 411 clips extracted from the given training set, leaving the remaining 20 clips as a test set. The best trial was selected from 5 different data partitions and 3 trials of the same model, each with randomized initial weights for each time scale. The predictions of the selected models were then averaged as the final output. In addition, an Extreme Learning Machine (ELM) method was trained for fusion, also averaging the outputs to produce the final emotion prediction. Finally, they also report the results of an SVR approach, here a global SVR that considered $6,373$ global song-level features extracted with OpenSMILE and a local SVR with 130 segment-level features, whose means and standard deviation were calculated with a 1s window and 0.5s shift, were added to form the final emotion prediction for each 0.5s clip. The results show that the proposed methods are significantly better than the baseline system, validating their multi-scale approach.

4 | Emotional Dynamics Models

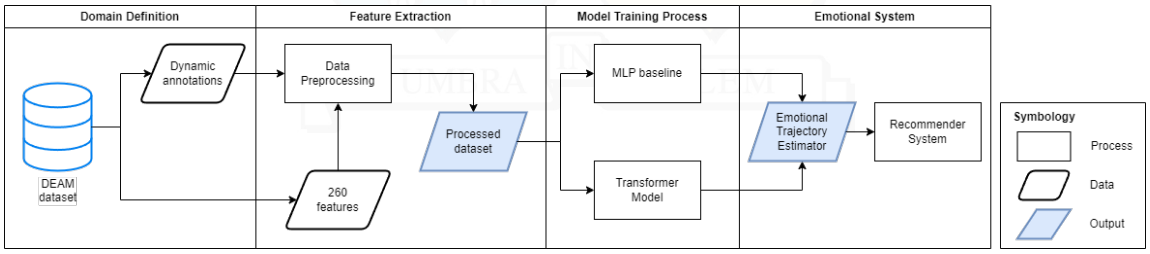


Figure 4.1: Flowchart depicting the implementation of the typical MER system steps in this work and the proposed Emotional System.

This chapter discusses the implementation of typical MER system steps, outlined in Subsection 2.1.2 and illustrated in Figure 4.1. The process involved in the data preprocessing, design, and development of selected techniques for obtaining an emotional trajectory estimator model are presented. In Data Preprocessing, considerations taken to filter and select data for defining training, test, and evaluation dataframes are outlined. Inspired by the works of Medina *et al.* (MLP baseline) [13] and Huang *et al.* (Transformer model) [14], the development process of the implemented models is detailed. Figures and tables are included to provide insights into the training steps and model selection process. The choice of the MLP model serves to validate the data preprocessing steps owing to its simplicity. In Section 4.2, the analysis performed on the obtained dataframes and the validation of the proposed metric criteria are presented. This includes an exploratory analysis of the model architecture parameters and plots depicting the prediction capabilities of each model configuration. Following the methodology of the previous section, Section 4.3 outlines the selection process of a Transformer-based model.

The objective of these models is to obtain an emotional trajectory estimator capable of predicting emotional transitions occurring over a piece of music, facilitating the detection of emotional flow over time. For this purpose, a model achieving low RMSE and high CCC metric values is desired, as these metrics are bound to the evaluation of the model's performance at a song level. This data can then be utilized to characterize each musical piece, detailing its emotional content and behavior. Such information facilitates the development of a system focused on the automatic generation of playlists based on emotional recommendations or recommending songs with similar emotional behavior.

4.1 Data Preprocessing

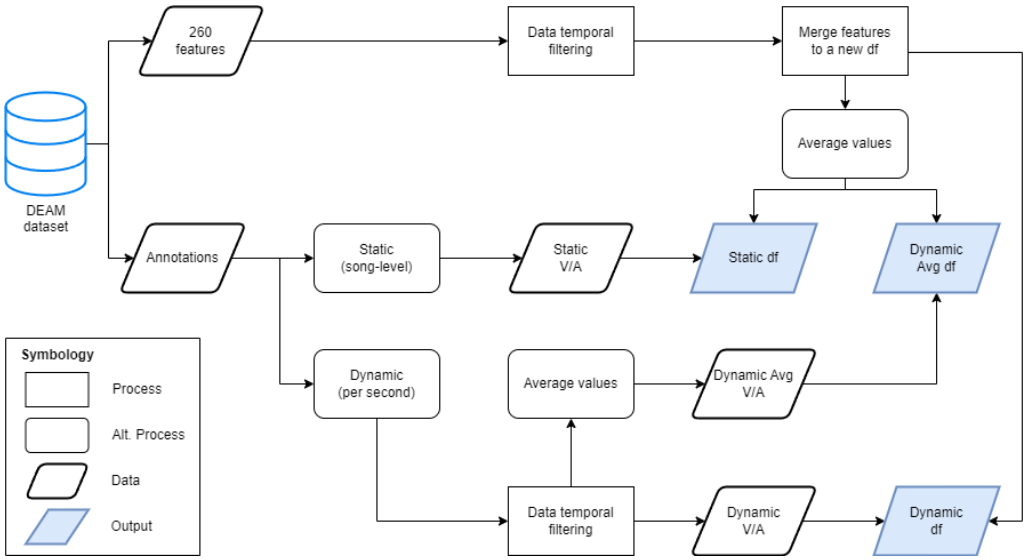


Figure 4.2: The flowchart details the steps of the Data Preprocessing Process using the values from the DEAM dataset.

The data available for download on the official DEAM dataset website [113] includes 1,802 audio files, along with song-level metadata and the features extracted from each song. The available annotation files comprise the dynamic and static (song level) annotations per each rater and averaged per song, obtained throughout the different editions of the MediaEval challenge (2013-2015). The flowchart in Figure 4.2 outlines the steps taken during the preprocessing of the DEAM dataset.

Adhering to the notes of Aljanaki *et al.* [70, 107], the initial 15 seconds of each song were excluded due to the stated instability and unreliability of the annotation process.

This exclusion allows annotators some habituation time before starting the record annotation process. Next, for the $1,744$ song excerpts, a time interval of 28.5 seconds ($t \in [15,000; 43,500][ms]$) containing non-zero dynamic annotation data values was extracted. Based on the available data, three dataframes are defined:

- *Static dataframe*: The static V/A annotations are paired by `song_id`, associating them with the average values of the 260 features available over time for each song. However, since these annotations are given on a 1-9 point scale, it was necessary to transform them onto a new scale ranging between -1 and 1 for each emotional dimension. Data: $1,802$ rows (songs) with 263 columns (`song_id, arousal, valence, 260 features`).

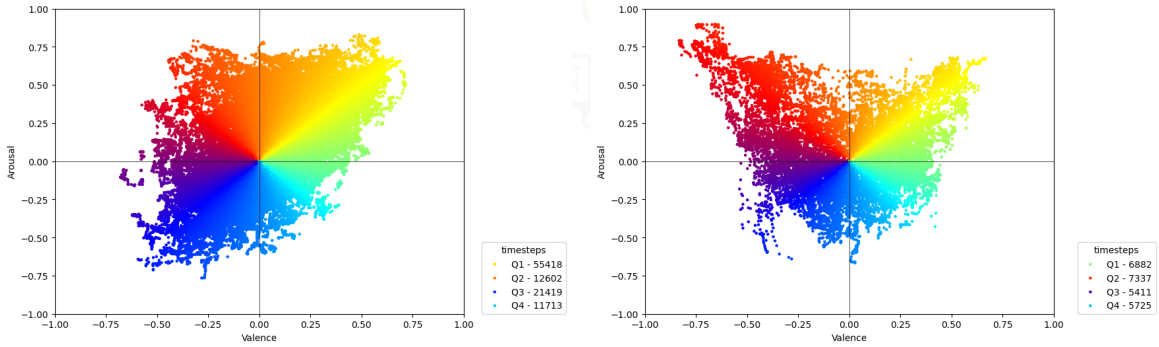
- *Dynamic Averaged dataframe*: The averaged values of dynamic annotations over time for each song are paired with the average feature values, mirroring the approach used for the static dataframe. Data: $1,802$ rows (songs) with 263 columns (`song_id, arousal, valence, 260 features`).

- *Dynamic dataframe*: Each (`song_id, timestamp`) pair was associated with the corresponding averaged arousal and valence dynamical annotation values across annotators. This file is characterized by its structure and serves as the main dataframe for the thesis. Data: $126,466$ rows (songs timesteps) with 264 columns (`song_id, timestep, arousal, valence, 260 features`).

For each dataframe, the available pool of songs was divided into two subsets: development and evaluation. The evaluation set comprised $25,314$ timesteps from the 58 full songs with `song_id` > 2000. Subsequently, the development set of each dataframe was split using an 80/20 ratio into training and test subsets, based on a randomized distribution of the available $1,744$ `song_id` values. To maintain consistency across the dataframes, the obtained `song_id` values were utilized to ensure the same song distribution. As a result, the training set comprised $80,794$ timesteps ($1,393$ songs), while the test set contained $20,358$ timesteps (351 songs).

For each dataframe, the training set was standardized using sklearn's `StandardScaler`.

Afterward, the fitted scaler was used to transform the testing and evaluation sets. Following this, a copy of the standardized training set underwent Principal Component Analysis (PCA) with 95% variance retention. This resulted in 60 components for the static and dynamic average dataframes, and 84 components for the dynamic dataframe. Figure 4.3 illustrates the quadrant distribution of timestep values in the Valence/Arousal plane, for the development and evaluation sets of the dynamic dataframe approach. These obtained dataframes are used as the basis for the various model configurations tested in this work, where the same configuration is trained for each emotional dimension separately.



(a) Development set **(b)** Evaluation set

Figure 4.3: DEAM V/A dynamic values distribution per timestamp with RGB color model based in Dharmapriya *et al.* work [11]. Red and orange colors represent high arousal values, and blue and green represent low arousal values, a convention also used by Coutinho *et al.* [12].

4.2 MLP Baseline

4.2.1 Model Definition

In their 2019 work, Medina *et al.* [13] introduced an MLP-based Emotion Prediction System, presenting its design process and the model configuration parameters evaluated across various experiments for a prediction system for automatic MER. Figure 4.4 illustrates the steps taken in developing the MLP model, with the *Model Defining* step involving the consideration of configuration parameters for different experiments:

- **Early stop:** The model utilized either an Early stop mechanism or underwent conventional training.
- **PCA:** with or without PCA applied on the training set.

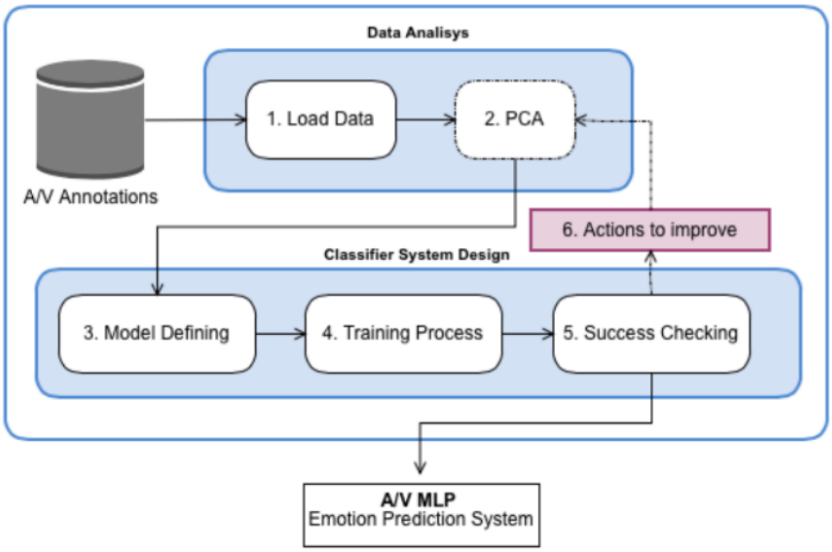


Figure 4.4: The principal phases of the Emotion Prediction System presented by Medina *et al.* [13].

- **Learning Rate:** {0.001, 0.010, 0.020, 0.030, 0.040, 0.050, 0.060, 0.070}.
- **Hidden Layers:** {1,2}.
- **Neurons:** {64, 128}.
- **Epochs:** 500.

Three distinct study cases were derived from each primary combination of configurations: early stop and no PCA (Exp. #1), conventional stop and no PCA (Exp. #2), and conventional stop and PCA (Exp. #3). Within each experimental setting, the internal configuration of the model architecture (choosing between 1 or 2 hidden layers with 64 or 128 neurons each) was systematically varied. The MLP models were implemented following a parameter exploration process akin to that presented by Medina *et al.*, testing different learning rate values for each experimental architecture and setting up to 500 epochs for the conventional stop approaches.

The nomenclature used in this work to reference results obtained from the implemented MLP models should be interpreted as follows: $\{A|V\}\_\{Exp\#\}\_\{HL\}\_\{Neurons\}\_\{LR\}$, where HL represents the number of hidden layers used, and LR denotes the learning rate value.

4.2.2 Implementation

This subsection presents the configuration of the models with the best RMSE scores for each approach, along with plots displaying estimations over the test and evaluation dataframes. The MLP model architecture was implemented as presented by Medina *et al.* with modifications limited to the architectural and configuration parameters outlined in the previous subsection. Three experiments were conducted based on each dataframe obtained in the Model Definition step. The simplicity of this architecture design allowed the use of the implemented model for comparison within works in the research area and the validation of the preprocessing analysis steps performed on the dataset.

The experiments encompass the same exploratory analysis conducted for the architecture model and involve variations in parameter configurations applied to the available dataframes. Tables displayed in Subsubsections 4.2.2.1 (Static), 4.2.2.2 (Dynamic Avg.), and 4.2.2.3 (Dynamic) present the Top 5 models for both test and evaluation sets, along with plots showcasing the predictive behavior of the Top 1 model for visual comparison. It's crucial to note that only RMSE is reported for the static and dynamic averaged approaches. This is due to the fact that, in these cases, only one emotional value is obtained for each song, leading to RMSE and MAE metrics exhibiting identical behaviors when averaged across songs. However, this distinction does not apply to the dynamic approach, where each song has multiple timesteps to compute the metrics. As a result, the presented standard deviation values reflect the deviation of song-level metrics compared to the average values of the full-song set.

4.2.2.1 Static

The lowest RMSE metric values obtained for the *static dataframe* approach in the testing and evaluation subsets are presented in Tables 4.1 and 4.2 respectively. Additionally, Table 4.3 highlights the top scores for each experiment, with their corresponding predictions displayed in Figure 4.5. The obtained metric values indicate a decrease in prediction accuracy for the valence dimension, as illustrated in the evaluation subfigures of Fig. 4.5. Conversely, the obtained arousal metrics suggest improved model performance.

Testing set

Table 4.1: Top 5 models with lowest RMSE scores implemented over the static dataframe testing set.

Exp #	Valence		Arousal	
	Model configuration	RMSE	Model configuration	RMSE
1	V_Exp1_2_128_0.010	0.170±0.128	A_Exp1_2_128_0.010	0.176±0.138
	V_Exp1_2_128_0.020	0.174±0.131	A_Exp1_1_128_0.010	0.181±0.152
	V_Exp1_2_128_0.030	0.176±0.132	A_Exp1_2_128_0.020	0.187±0.156
	V_Exp1_2_64_0.050	0.179±0.135	A_Exp1_1_128_0.070	0.190±0.161
	V_Exp1_2_64_0.040	0.181±0.139	A_Exp1_1_128_0.020	0.191±0.171
2	V_Exp2_2_128_0.020	0.173±0.146	A_Exp2_2_128_0.020	0.196±0.166
	V_Exp2_2_64_0.050	0.177±0.150	A_Exp2_2_128_0.050	0.197±0.154
	V_Exp2_2_64_0.010	0.182±0.142	A_Exp2_2_128_0.030	0.197±0.159
	V_Exp2_2_128_0.030	0.190±0.150	A_Exp2_2_128_0.010	0.200±0.166
	V_Exp2_1_128_0.070	0.195±0.154	A_Exp2_2_128_0.040	0.201±0.150
3	V_Exp3_2_128_0.010	0.180±0.132	A_Exp3_2_128_0.010	0.196±0.160
	V_Exp3_1_128_0.010	0.190±0.149	A_Exp3_1_128_0.010	0.203±0.177
	V_Exp3_2_64_0.050	0.190±0.154	A_Exp3_2_64_0.020	0.206±0.169
	V_Exp3_1_64_0.060	0.191±0.142	A_Exp3_2_64_0.030	0.212±0.175
	V_Exp3_2_128_0.040	0.197±0.152	A_Exp3_2_128_0.020	0.213±0.186

Evaluation set

Table 4.2: Top 5 models with lowest RMSE scores implemented over the static dataframe evaluation set.

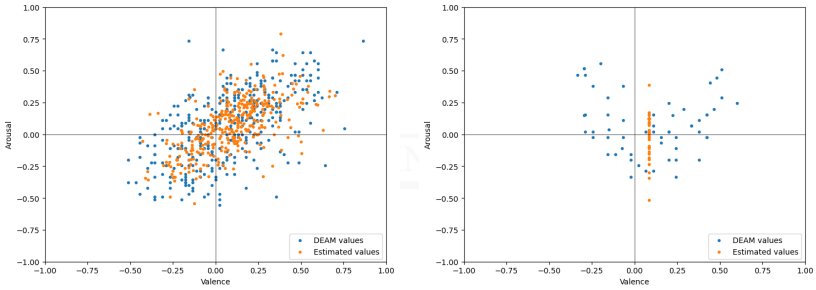
Exp #	Valence		Arousal	
	Model configuration	RMSE	Model configuration	RMSE
1	V_Exp1_2_64_0.070	0.229±0.130	A_Exp1_2_128_0.040	0.171±0.137
	V_Exp1_2_128_0.030	0.243±0.147	A_Exp1_2_128_0.030	0.187±0.138
	V_Exp1_2_128_0.010	0.246±0.141	A_Exp1_2_128_0.010	0.189±0.163
	V_Exp1_2_128_0.020	0.263±0.189	A_Exp1_2_64_0.060	0.206±0.206
	V_Exp1_2_128_0.060	0.275±0.182	A_Exp1_2_128_0.050	0.242±0.183
2	V_Exp2_2_64_0.060	0.229±0.131	A_Exp2_2_64_0.030	0.183±0.139
	V_Exp2_2_64_0.070	0.229±0.131	A_Exp2_2_64_0.070	0.185±0.141
	V_Exp2_2_64_0.040	0.229±0.132	A_Exp2_2_64_0.060	0.186±0.140
	V_Exp2_2_128_0.020	0.249±0.168	A_Exp2_2_64_0.050	0.199±0.214
	V_Exp2_2_128_0.030	0.261±0.183	A_Exp2_2_64_0.050	0.200±0.133
3	V_Exp3_2_128_0.070	0.228±0.134	A_Exp3_2_64_0.070	0.186±0.140
	V_Exp3_2_128_0.060	0.228±0.132	A_Exp3_2_128_0.050	0.186±0.139
	V_Exp3_2_128_0.050	0.229±0.131	A_Exp3_2_128_0.060	0.187±0.138
	V_Exp3_2_64_0.070	0.229±0.131	A_Exp3_2_128_0.070	0.189±0.136
	V_Exp3_2_64_0.060	0.274±0.183	A_Exp3_1_64_0.040	0.262±0.228

Top models

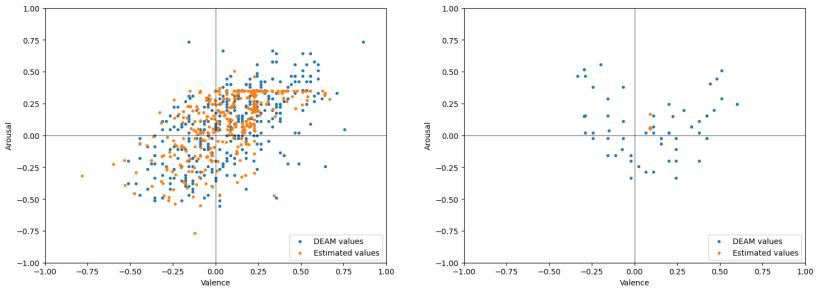
Table 4.3: Top models implemented over the evaluation set of the static dataframe.

Experiments Settings			Valence	Arousal
Exp #	Mode	PCA	RMSE	RMSE
1	Early stop	No	0.229±0.130	0.171±0.137
2	Conventional	No	0.229±0.131	0.183±0.139
3	Conventional	Yes	0.228±0.134	0.186±0.140

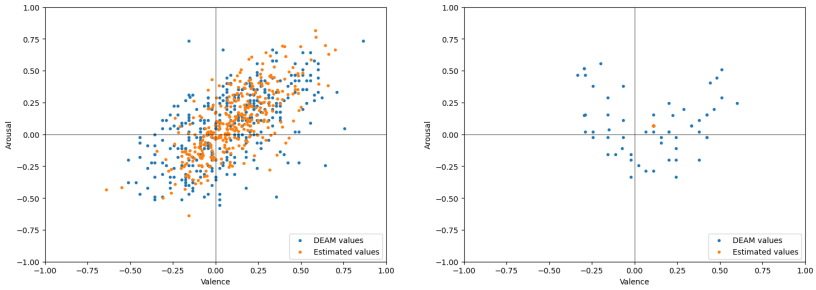
Top models predictions



(a) Exp 1, *test* **(b)** Exp 1, *eval*



(c) Exp 2, *test* **(d)** Exp 2, *eval*



(e) Exp 3, *test* **(f)** Exp 3, *eval*

Figure 4.5: Visualization of the predictions obtained with the Top model for DEAM static dataframe approach.

4.2.2.2 Dynamic averaged

The lowest RMSE metric values obtained for the *dynamic avgeraged dataframe* approach in the testing and evaluation subsets are presented in Tables 4.4 and 4.5 respectively. Additionally, Table 4.6 highlights the top scores for each experiment, with their corresponding predictions displayed in Figure 4.6. The obtained metric values not only indicate a decrease in prediction accuracy for the valence dimension but also for the arousal dimension, as illustrated in the evaluation subfigures of Fig. 4.6, similar to the previous case. Although the top RMSE scores are slightly higher than those in the static approach, this difference can be attributed to the changes in the arousal and valence reference values for the same `song_id`, given the model's consistent behavior in both cases.

Testing set

Table 4.4: Top 5 models with lowest RMSE scores implemented over the dynamic averaged dataframe testing set.

Exp #	Valence		Arousal	
	Model configuration	RMSE	Model configuration	RMSE
1	V_Exp1_2_128_0.060	0.168±0.129	A_Exp1_2_64_0.010	0.150±0.131
	V_Exp1_2_128_0.010	0.168±0.120	A_Exp1_2_64_0.020	0.152±0.130
	V_Exp1_2_128_0.050	0.169±0.128	A_Exp1_2_128_0.040	0.161±0.132
	V_Exp1_2_64_0.020	0.169±0.130	A_Exp1_1_128_0.040	0.163±0.127
	V_Exp1_2_128_0.030	0.170±0.127	A_Exp1_1_128_0.020	0.164±0.129
2	V_Exp2_2_64_0.020	0.164±0.127	A_Exp2_1_64_0.020	0.164±0.136
	V_Exp2_2_64_0.010	0.176±0.139	A_Exp2_2_64_0.040	0.169±0.134
	V_Exp2_2_64_0.050	0.179±0.138	A_Exp2_2_64_0.030	0.171±0.139
	V_Exp2_2_128_0.020	0.180±0.136	A_Exp2_2_64_0.050	0.171±0.128
	V_Exp2_1_128_0.040	0.181±0.137	A_Exp2_1_128_0.010	0.172±0.139
3	V_Exp3_2_128_0.010	0.181±0.134	A_Exp3_1_128_0.010	0.156±0.133
	V_Exp3_2_128_0.050	0.187±0.130	A_Exp3_2_128_0.010	0.160±0.131
	V_Exp3_2_128_0.060	0.187±0.130	A_Exp3_2_64_0.020	0.168±0.148
	V_Exp3_2_64_0.030	0.187±0.140	A_Exp3_1_64_0.030	0.171±0.137
	V_Exp3_2_128_0.070	0.188±0.130	A_Exp3_2_64_0.040	0.171±0.142

Evaluation set

Table 4.5: Top 5 models with lowest RMSE scores implemented over the dynamic averaged dataframe evaluation set.

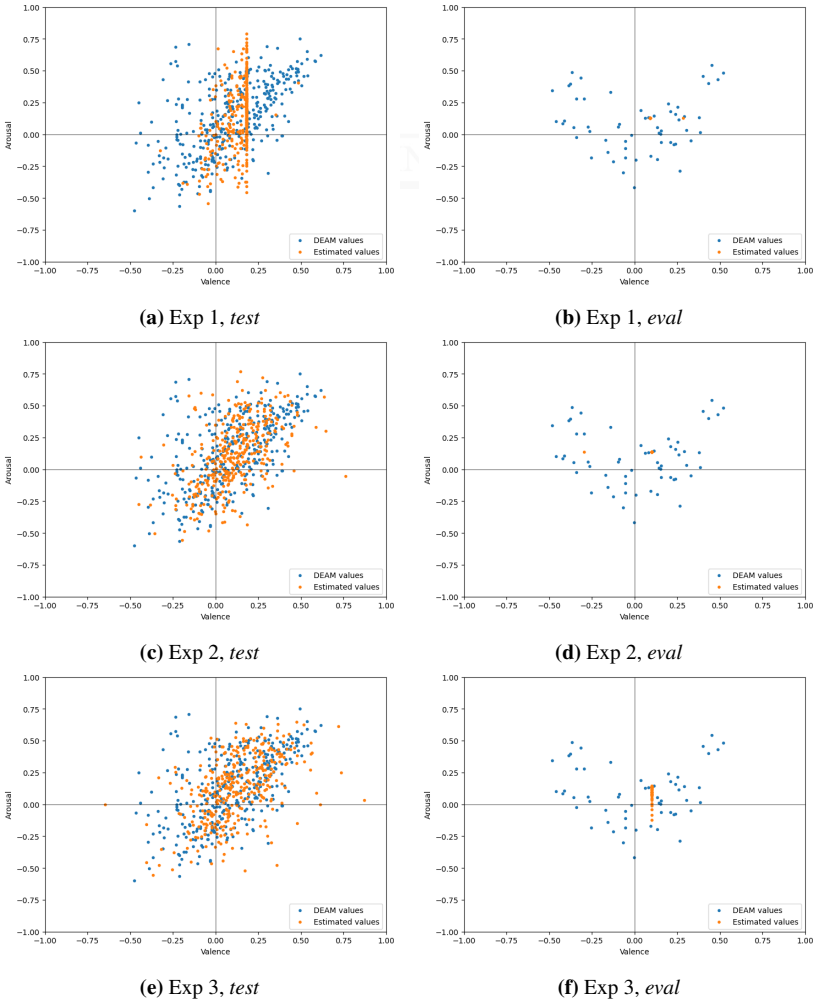
Exp #	Valence		Arousal	
	Model configuration	RMSE	Model configuration	RMSE
1	V_Exp1_2_64_0.060	0.234±0.162	A_Exp1_2_64_0.070	0.181±0.132
	V_Exp1_2_128_0.030	0.235±0.175	A_Exp1_1_128_0.020	0.182±0.173
	V_Exp1_2_64_0.070	0.236±0.158	A_Exp1_2_128_0.050	0.183±0.132
	V_Exp1_2_128_0.050	0.239±0.179	A_Exp1_2_128_0.060	0.184±0.132
	V_Exp1_2_128_0.020	0.245±0.169	A_Exp1_2_128_0.020	0.202±0.238
2	V_Exp2_2_64_0.060	0.231±0.158	A_Exp2_2_128_0.060	0.183±0.132
	V_Exp2_2_64_0.040	0.237±0.163	A_Exp2_2_64_0.070	0.184±0.132
	V_Exp2_2_128_0.040	0.237±0.163	A_Exp2_2_64_0.060	0.184±0.132
	V_Exp2_2_64_0.070	0.237±0.163	A_Exp2_2_128_0.001	0.239±0.271
	V_Exp2_2_128_0.060	0.240±0.179	A_Exp2_2_64_0.030	0.282±0.216
3	V_Exp3_2_64_0.070	0.238±0.164	A_Exp3_2_64_0.050	0.174±0.142
	V_Exp3_2_128_0.050	0.238±0.164	A_Exp3_2_64_0.070	0.184±0.132
	V_Exp3_2_128_0.060	0.238±0.167	A_Exp3_2_128_0.060	0.185±0.131
	V_Exp3_2_128_0.070	0.239±0.170	A_Exp3_2_128_0.070	0.188±0.131
	V_Exp3_2_64_0.060	0.241±0.164	A_Exp3_2_128_0.040	0.230±0.183

Top models

Table 4.6: Top models implemented over the evaluation set of the dynamic averaged dataframe.

Experiments Settings			Valence	Arousal
Exp #	Mode	PCA	RMSE	RMSE
1	Early stop	No	0.234±0.162	0.181±0.132
2	Conventional	No	0.231±0.158	0.183±0.132
3	Conventional	Yes	0.238±0.164	0.174±0.142

Top models predictions



(a) Exp 1, *test* **(b)** Exp 1, *eval*

(c) Exp 2, *test* **(d)** Exp 2, *eval*

(e) Exp 3, *test* **(f)** Exp 3, *eval*

Figure 4.6: Visualization of the predictions obtained with the Top model for DEAM dynamic averaged dataframe approach.

4.2.2.3 Dynamic

This experiment acts as the baseline for the rest of the thesis, for which the MAE and CCC metrics are also reported. To keep consistency in the comparison with the previous implementations, Tables 4.7 and 4.8 showcase the testing and evaluation models, respectively, with top RMSE scores. Additionally, Table 4.9 highlights the models with top RMSE scores for each experiment, with their corresponding predictions displayed in Figure 4.7. The obtained RMSE values indicate a significant decrease in prediction accuracy in both dimensions, as illustrated in the evaluation subfigures of Fig. 4.7. In this scenario, the CCC metric provides insight into the model's behavior, suggesting that the obtained model for the valence dimension lacks useful prediction capabilities in this context, which aligns with the observed model behavior.

Testing set

Table 4.7: Top 5 models with lowest RMSE scores implemented over the dynamic dataframe testing set.

Exp #	Valence				Arousal			
	Model configuration	RMSE	MAE	CCC	Model configuration	RMSE	MAE	CCC
1	V_Exp1_2_128_0.030	0.177±0.111	0.169±0.114	0.008±0.088	A_Exp1_1_64_0.001	0.175±0.092	0.154±0.092	0.031±0.151
	V_Exp1_2_64_0.030	0.178±0.099	0.163±0.100	0.012±0.100	A_Exp1_1_64_0.030	0.176±0.098	0.156±0.099	0.023±0.145
	V_Exp1_1_64_0.020	0.179±0.091	0.162±0.093	0.012±0.097	A_Exp1_2_128_0.020	0.176±0.098	0.155±0.099	0.028±0.149
	V_Exp1_2_64_0.040	0.180±0.103	0.167±0.105	0.008±0.095	A_Exp1_2_64_0.030	0.176±0.096	0.154±0.097	0.024±0.145
	V_Exp1_2_64_0.020	0.180±0.093	0.163±0.095	0.015±0.098	A_Exp1_2_64_0.020	0.177±0.088	0.155±0.087	0.026±0.136
2	V_Exp2_1_64_0.001	0.177±0.092	0.160±0.094	0.017±0.099	A_Exp2_2_64_0.020	0.178±0.099	0.157±0.099	0.028±0.140
	V_Exp2_2_128_0.010	0.182±0.095	0.164±0.097	0.011±0.093	A_Exp2_2_128_0.020	0.179±0.102	0.160±0.102	0.019±0.131
	V_Exp2_2_64_0.020	0.183±0.094	0.165±0.095	0.012±0.096	A_Exp2_2_64_0.010	0.179±0.098	0.158±0.098	0.029±0.141
	V_Exp2_2_64_0.010	0.192±0.097	0.169±0.097	0.018±0.083	A_Exp2_2_128_0.001	0.188±0.096	0.163±0.094	0.031±0.130
	V_Exp2_2_128_0.030	0.193±0.126	0.189±0.128	0.000±0.000	A_Exp2_2_128_0.010	0.190±0.089	0.163±0.088	0.026±0.128
3	V_Exp3_1_64_0.020	0.176±0.100	0.162±0.102	0.007±0.111	A_Exp3_1_128_0.010	0.178±0.099	0.157±0.099	0.028±0.145
	V_Exp3_1_64_0.030	0.181±0.110	0.167±0.108	0.009±0.103	A_Exp3_1_64_0.020	0.181±0.096	0.160±0.095	0.018±0.147
	V_Exp3_1_64_0.040	0.181±0.105	0.166±0.105	0.012±0.103	A_Exp3_2_128_0.030	0.181±0.094	0.160±0.094	0.020±0.137
	V_Exp3_2_64_0.020	0.189±0.094	0.169±0.094	0.017±0.093	A_Exp3_2_128_0.020	0.183±0.090	0.159±0.089	0.026±0.132
	V_Exp3_2_64_0.010	0.190±0.094	0.168±0.095	0.014±0.094	A_Exp3_1_64_0.010	0.183±0.093	0.160±0.093	0.026±0.133

Evaluation set

Table 4.8: Top 5 models with lowest RMSE scores implemented over the dynamic dataframe evaluation set.

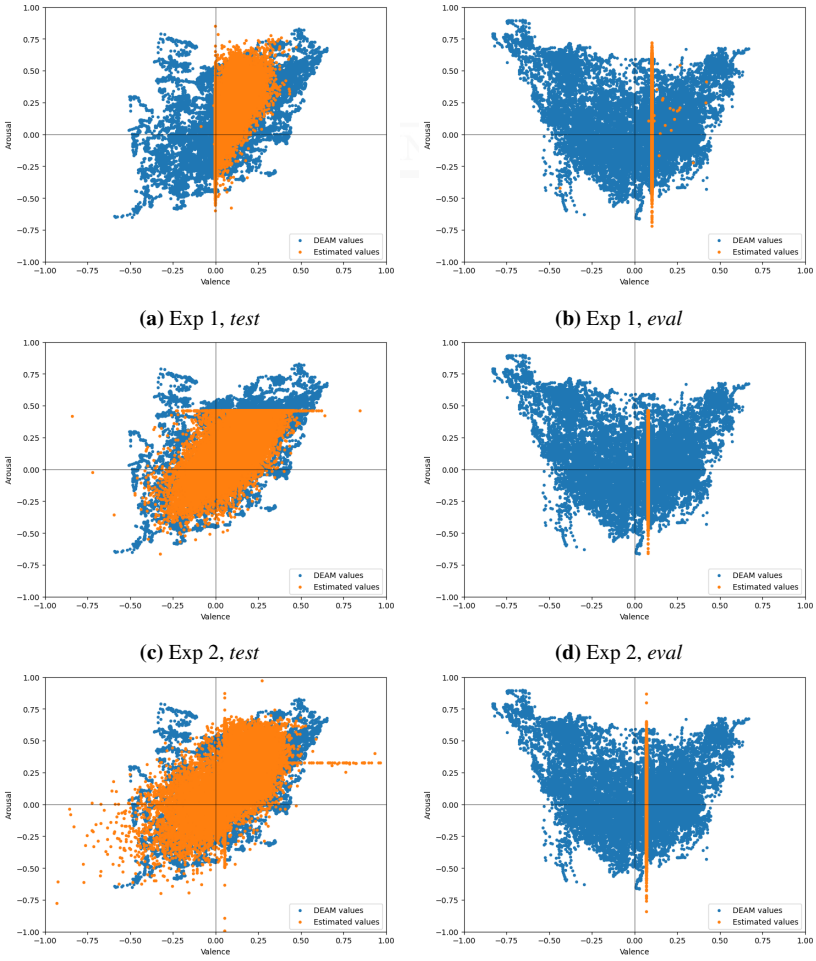
Exp #	Valence				Arousal			
	Model configuration	RMSE	MAE	CCC	Model configuration	RMSE	MAE	CCC
1	V_Exp1_2_128_0.070	0.267±0.152	0.249±0.152	0.000±0.001	A_Exp1_1_64_0.001	0.222±0.067	0.192±0.067	0.275±0.236
	V_Exp1_2_64_0.060	0.268±0.152	0.250±0.152	0.000±0.002	A_Exp1_2_64_0.030	0.222±0.062	0.191±0.063	0.271±0.237
	V_Exp1_2_128_0.040	0.268±0.152	0.250±0.152	0.000±0.000	A_Exp1_1_64_0.030	0.223±0.068	0.194±0.068	0.288±0.240
	V_Exp1_2_128_0.060	0.268±0.152	0.250±0.152	0.000±0.000	A_Exp1_1_64_0.060	0.223±0.070	0.194±0.072	0.257±0.222
	V_Exp1_2_64_0.070	0.268±0.153	0.250±0.153	0.001±0.006	A_Exp1_1_128_0.010	0.224±0.075	0.193±0.076	0.289±0.250
2	V_Exp2_2_128_0.020	0.265±0.144	0.247±0.144	0.000±0.000	A_Exp2_2_64_0.020	0.228±0.070	0.195±0.071	0.265±0.231
	V_Exp2_2_128_0.070	0.265±0.145	0.247±0.145	0.000±0.000	A_Exp2_2_64_0.010	0.230±0.067	0.198±0.068	0.235±0.224
	V_Exp2_2_128_0.050	0.266±0.146	0.248±0.146	0.000±0.000	A_Exp2_2_128_0.020	0.236±0.072	0.205±0.073	0.232±0.219
	V_Exp2_2_64_0.040	0.266±0.147	0.248±0.147	0.000±0.000	A_Exp2_2_128_0.001	0.236±0.064	0.201±0.063	0.210±0.229
	V_Exp2_2_64_0.030	0.266±0.149	0.248±0.149	0.000±0.000	A_Exp2_2_128_0.010	0.243±0.068	0.206±0.067	0.231±0.223
3	V_Exp3_2_128_0.070	0.265±0.142	0.247±0.142	0.000±0.000	A_Exp3_1_64_0.020	0.224±0.075	0.195±0.077	0.278±0.230
	V_Exp3_2_64_0.070	0.265±0.145	0.247±0.145	0.000±0.000	A_Exp3_1_64_0.010	0.226±0.072	0.196±0.073	0.272±0.214
	V_Exp3_2_64_0.040	0.266±0.149	0.248±0.149	0.000±0.000	A_Exp3_1_64_0.030	0.227±0.079	0.197±0.080	0.293±0.218
	V_Exp3_2_64_0.060	0.266±0.149	0.249±0.149	0.000±0.000	A_Exp3_2_64_0.010	0.230±0.070	0.198±0.071	0.258±0.235
	V_Exp3_2_64_0.050	0.267±0.152	0.250±0.152	0.000±0.000	A_Exp3_1_128_0.010	0.232±0.071	0.201±0.072	0.257±0.228

Top models

Table 4.9: Top models implemented over the evaluation set of the dynamic dataframe.

Experiments Settings			Valence			Arousal		
Exp #	Mode	PCA	RMSE	MAE	CCC	RMSE	MAE	CCC
1	Early stop	No	0.267±0.152	0.249±0.152	0.000±0.001	0.222±0.067	0.192±0.067	0.275±0.236
2	Conventional	No	0.265±0.144	0.247±0.144	0.000±0.000	0.228±0.070	0.195±0.071	0.265±0.231
3	Conventional	Yes	0.265±0.142	0.247±0.142	0.000±0.000	0.224±0.075	0.195±0.077	0.278±0.230

Top models predictions



(a) Exp 1, *test* **(b)** Exp 1, *eval*

(c) Exp 2, *test* **(d)** Exp 2, *eval*

(e) Exp 3, *test* **(f)** Exp 3, *eval*

Figure 4.7: Visualization of the predictions obtained with the Top model for DEAM dynamic dataframe approach.

4.2.3 Model Analysis

If the model evaluation relied solely on the averaged RMSE score across songs without considering the temporal behavior of these predictions, the recommendations might be even worse than suboptimal, as presented in the previous subsection. Figures 4.5 (Static), 4.6 (Dynamic Avg.), and 4.7 (Dynamic) depict the predictive behavior of the obtained top RMSE models for each dataframe. In both, static (Table 4.2) and dynamic averaged (Table 4.5) approaches, the models tend to obtain better metrics for the architectures with two hidden-layers. It is also noticeable that there is already a difference between the prediction metric values for both dimensions, aligning with the known fact that the valence dimension is more complex to estimate and predict than the arousal dimension.

In the dynamic case (Table 4.8), it immediately comes to sight the values of the CCC metric for top RMSE Valence models, as all of them tend to be zero. This added to the high standard deviation values in the RMSE metric in comparison with the arousal dimension, indicates that, regardless of the values being predicted somewhat near the expected value for some songs, these values do not correlate with the expected values. This can be easily explained visually in the subfigures related to the evaluation set in Fig. 4.7. However, along with achieving somewhat acceptable prediction metric values for the testing set, the plots displayed visual similarities between predictions and the reference DEAM values, partially validating the previous preprocessing steps. The proposed modification to the Top model selection criteria aims to establish a baseline MLP model that generates meaningful estimations. Instead of relying solely on the RMSE score, an averaged CCC metric across songs is suggested, as it closely aligns with the behavior of the estimation values within a particular song. This approach is implemented exclusively for the Dynamic approach, aligning with the primary focus of this research on developing a Dynamic MER-based MRS.

4.2.3.1 Top CCC models

Aiming to obtain models with better prediction capabilities, the top CCC models are selected. Tables 4.10 and 4.11 showcase the testing and evaluation models respectively. Additionally, Table 4.12 highlights the models with top CCC scores for each experiment, with their corresponding predictions displayed in Figure 4.8.

Testing set

Table 4.10: Top 5 models with highest CCC scores implemented over the dynamic dataframe testing set.

Exp #	Valence				Arousal			
	Model configuration	RMSE	MAE	CCC	Model configuration	RMSE	MAE	CCC
1	V_Exp1_2_64_0.010	0.189±0.091	0.167±0.092	0.018±0.099	A_Exp1_1_64_0.001	0.175±0.092	0.154±0.092	0.031±0.151
	V_Exp1_2_128_0.001	0.186±0.086	0.164±0.087	0.016±0.087	A_Exp1_2_128_0.030	0.180±0.097	0.158±0.097	0.030±0.147
	V_Exp1_1_64_0.001	0.182±0.093	0.164±0.094	0.015±0.094	A_Exp1_2_128_0.010	0.182±0.087	0.158±0.087	0.029±0.131
	V_Exp1_2_64_0.020	0.180±0.093	0.163±0.095	0.015±0.098	A_Exp1_1_64_0.001	0.185±0.091	0.161±0.091	0.028±0.134
	V_Exp1_1_64_0.060	0.198±0.096	0.177±0.097	0.014±0.092	A_Exp1_2_128_0.020	0.176±0.098	0.155±0.099	0.028±0.149
2	V_Exp2_2_64_0.010	0.192±0.097	0.169±0.097	0.018±0.083	A_Exp2_2_128_0.001	0.188±0.096	0.163±0.094	0.031±0.130
	V_Exp2_1_64_0.001	0.177±0.092	0.160±0.094	0.017±0.099	A_Exp2_2_64_0.010	0.179±0.098	0.158±0.098	0.029±0.141
	V_Exp2_2_64_0.020	0.183±0.094	0.165±0.095	0.012±0.096	A_Exp2_2_64_0.020	0.178±0.099	0.157±0.099	0.028±0.140
	V_Exp2_2_128_0.010	0.182±0.095	0.164±0.097	0.011±0.093	A_Exp2_2_128_0.010	0.190±0.089	0.163±0.088	0.026±0.128
	V_Exp2_2_64_0.001	0.208±0.09	0.178±0.089	0.010±0.076	A_Exp2_1_64_0.001	0.208±0.100	0.176±0.094	0.023±0.115
3	V_Exp3_2_64_0.020	0.189±0.094	0.169±0.094	0.017±0.093	A_Exp3_2_64_0.020	0.185±0.097	0.162±0.097	0.034±0.136
	V_Exp3_2_64_0.030	0.197±0.098	0.175±0.100	0.015±0.095	A_Exp3_2_128_0.010	0.189±0.090	0.162±0.088	0.033±0.131
	V_Exp3_2_128_0.010	0.190±0.094	0.168±0.095	0.014±0.094	A_Exp3_1_128_0.010	0.178±0.099	0.157±0.099	0.028±0.145
	V_Exp3_2_64_0.010	0.191±0.095	0.169±0.096	0.013±0.090	A_Exp3_2_128_0.020	0.183±0.090	0.159±0.089	0.026±0.132
	V_Exp3_2_128_0.050	0.194±0.098	0.174±0.098	0.013±0.087	A_Exp3_2_64_0.010	0.183±0.093	0.160±0.093	0.026±0.133

Evaluation set

Table 4.11: Top 5 models with highest CCC scores implemented over the dynamic dataframe evaluation set.

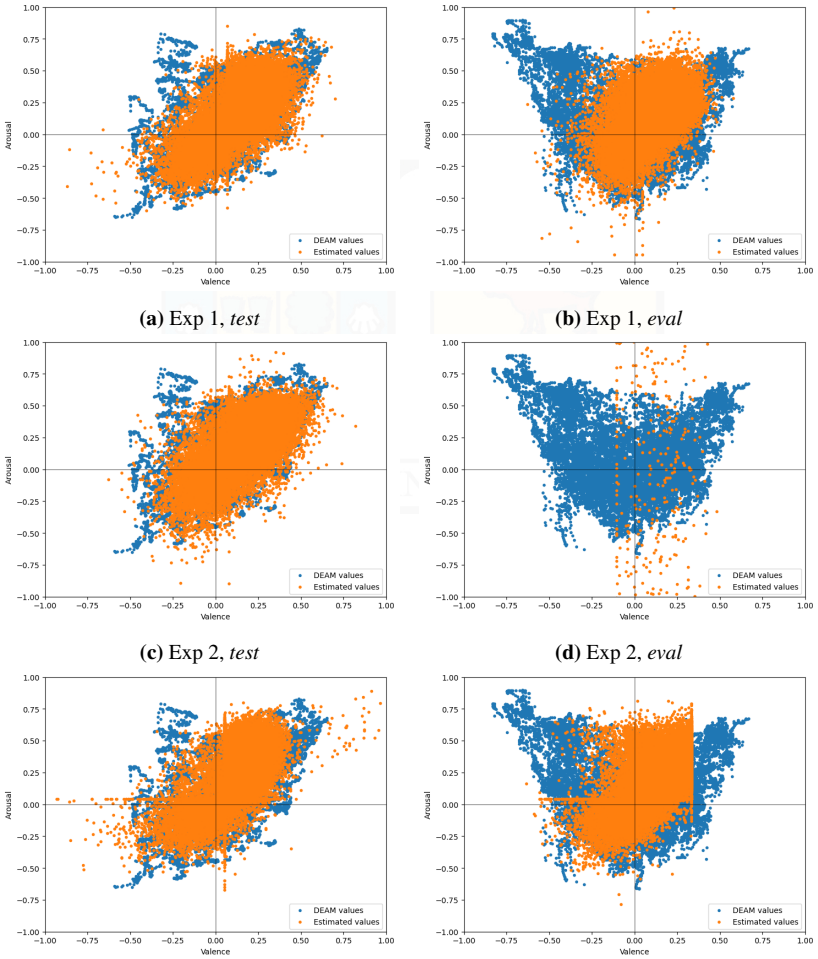
Exp #	Valence				Arousal			
	Model configuration	RMSE	MAE	CCC	Model configuration	RMSE	MAE	CCC
1	V_Exp1_2_64_0.020	0.276±0.151	0.246±0.148	0.058±0.108	A_Exp1_1_128_0.001	0.224±0.075	0.193±0.076	0.289±0.250
	V_Exp1_2_64_0.030	0.286±0.171	0.256±0.167	0.056±0.136	A_Exp1_1_64_0.030	0.223±0.068	0.194±0.068	0.288±0.240
	V_Exp1_2_128_0.020	0.284±0.160	0.253±0.156	0.056±0.130	A_Exp1_2_64_0.050	0.234±0.070	0.201±0.070	0.277±0.227
	V_Exp1_2_64_0.050	0.284±0.158	0.253±0.155	0.049±0.132	A_Exp1_1_64_0.001	0.222±0.067	0.192±0.067	0.275±0.236
	V_Exp1_2_64_0.001	0.293±0.158	0.259±0.154	0.046±0.110	A_Exp1_2_128_0.030	0.233±0.072	0.200±0.073	0.275±0.237
2	V_Exp2_2_128_0.010	0.292±0.152	0.260±0.149	0.054±0.121	A_Exp2_1_128_0.030	0.236±0.083	0.201±0.081	0.308±0.228
	V_Exp2_2_128_0.050	0.293±0.135	0.258±0.135	0.052±0.124	A_Exp2_1_64_0.030	0.227±0.079	0.197±0.080	0.293±0.218
	V_Exp2_2_64_0.030	0.307±0.163	0.271±0.161	0.050±0.109	A_Exp2_1_64_0.020	0.224±0.075	0.195±0.077	0.278±0.230
	V_Exp2_1_64_0.030	0.282±0.162	0.254±0.161	0.046±0.141	A_Exp2_1_128_0.040	0.243±0.080	0.205±0.075	0.276±0.222
	V_Exp2_1_64_0.040	0.289±0.159	0.260±0.157	0.043±0.137	A_Exp2_1_64_0.020	0.226±0.072	0.196±0.073	0.272±0.214
3	V_Exp3_2_128_0.010	0.284±0.154	0.254±0.150	0.043±0.104	A_Exp3_2_64_0.020	0.228±0.070	0.195±0.071	0.265±0.231
	V_Exp3_2_64_0.020	0.278±0.145	0.247±0.141	0.042±0.101	A_Exp3_1_64_0.001	0.252±0.079	0.210±0.066	0.244±0.230
	V_Exp3_1_64_0.001	0.286±0.154	0.255±0.151	0.039±0.113	A_Exp3_2_64_0.010	0.230±0.067	0.198±0.068	0.235±0.224
	V_Exp3_2_64_0.010	0.304±0.172	0.268±0.166	0.038±0.115	A_Exp3_2_128_0.020	0.236±0.072	0.205±0.073	0.232±0.219
	V_Exp3_2_128_0.001	0.305±0.140	0.266±0.138	0.031±0.080	A_Exp3_2_128_0.010	0.243±0.068	0.206±0.067	0.231±0.223

Top models

Table 4.12: Top models implemented over the evaluation set of the dynamic dataframe.

Experiments Settings			Valence			Arousal		
Exp #	Mode	PCA	RMSE	MAE	CCC	RMSE	MAE	CCC
1	Early stop	No	0.276±0.151	0.246±0.148	0.058±0.108	0.224±0.075	0.193±0.076	0.289±0.250
2	Conventional	No	0.292±0.152	0.260±0.149	0.054±0.121	0.236±0.083	0.201±0.081	0.308±0.228
3	Conventional	Yes	0.284±0.154	0.254±0.150	0.043±0.104	0.228±0.070	0.195±0.071	0.265±0.231

Top models prediction



(a) Exp 1, *test* **(b)** Exp 1, *eval*

(c) Exp 2, *test* **(d)** Exp 2, *eval*

(e) Exp 3, *test* **(f)** Exp 3, *eval*

Figure 4.8: Visualization of the predictions obtained with top CCC models for DEAM dynamic dataframe approach.

4.2.3.2 Metrics comparison

The Figures 4.9 (Valence) and 4.10 (Arousal) depict the changes in metric values resulting from variations in the selected top model criteria. In both cases, RMSE (mean and std) values tend to increase, as do the MAE values, deviating from the typical goal of minimizing these measurements to achieve a better model. However, an increase in the CCC metric is observed alongside graphic prediction variation compared to the values obtained with the MLP models. Therefore, in these cases, the tradeoff between finding the minimum RMSE score metric and neglecting the CCC values could lead to significantly increased prediction errors, as visualized in the previous subsections, resulting in performance issues in the final

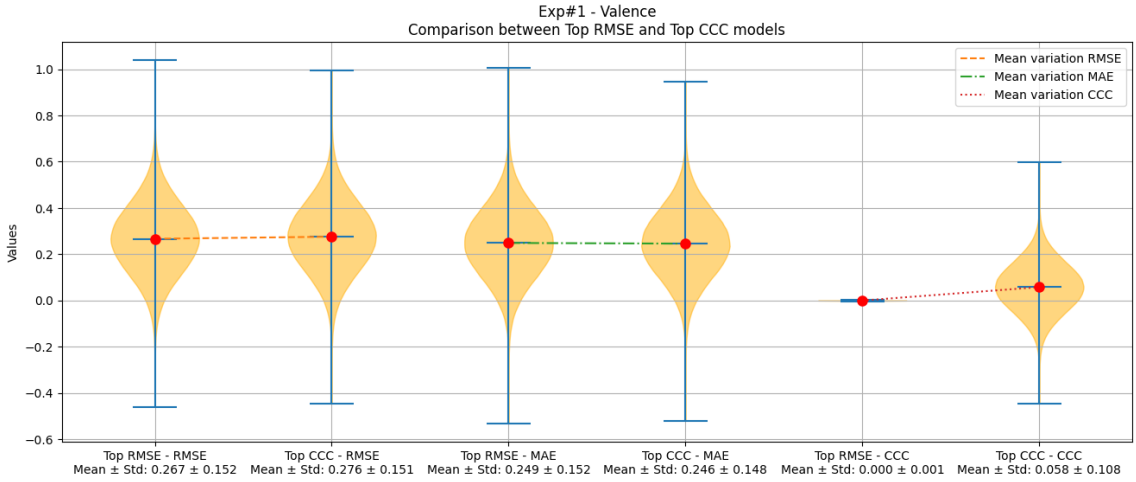
application.

While the predictions still fall short of the desired accuracy, the metric values obtained for the top RMSE models can be compared to the reported values of Medina *et al.* [13] (Table 4.13). In this comparison, arousal models outperform the reported values in both metrics, but there is a decline in performance in valence dimension prediction.

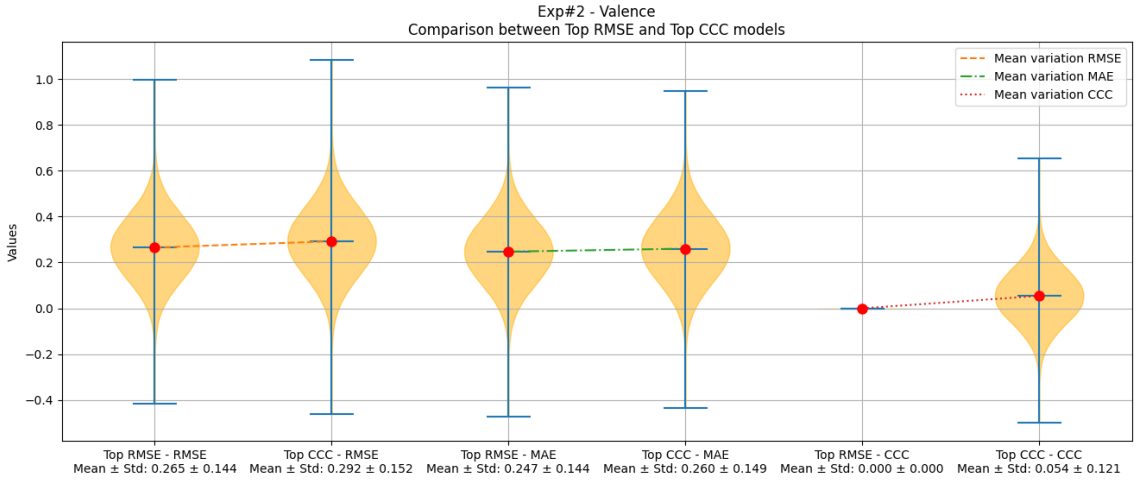
Medina reported the best model with RMSE scores of 0.24 for Arousal and 0.23 for Valence dimensions. In comparison, the best-reported models obtained from the RMSE top yielded scores of 0.222 ± 0.067 for Arousal (Exp #1) and 0.265 ± 0.142 for the Valence (Exp #3) dimension. Based on this, both data preprocessing and model implementation steps are considered to be successfully adapted from Medina *et al.* work, thus establishing the obtained models and predictions as baseline values for future comparisons in this work. Additionally, inclusion of CCC values provides an expanded set of metrics for comparison and validation of the predictions.

Table 4.13: Reported *best test scenarios* of Medina *et al.* work.

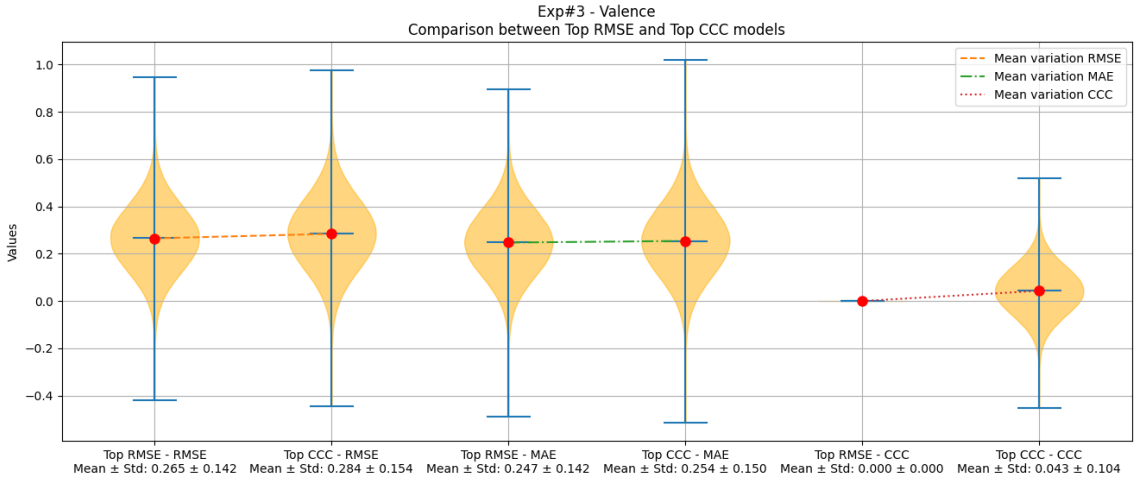
Experiments Settings			Valence		Arousal	
Exp #	Mode	PCA	RMSE	MAE	RMSE	MAE
1	Early stop	No	0.25	0.21	0.27	0.22
2	Conventional	No	0.25	0.20	0.28	0.23
3	Conventional	Yes	0.23	0.18	0.24	0.20



(a) Exp #1

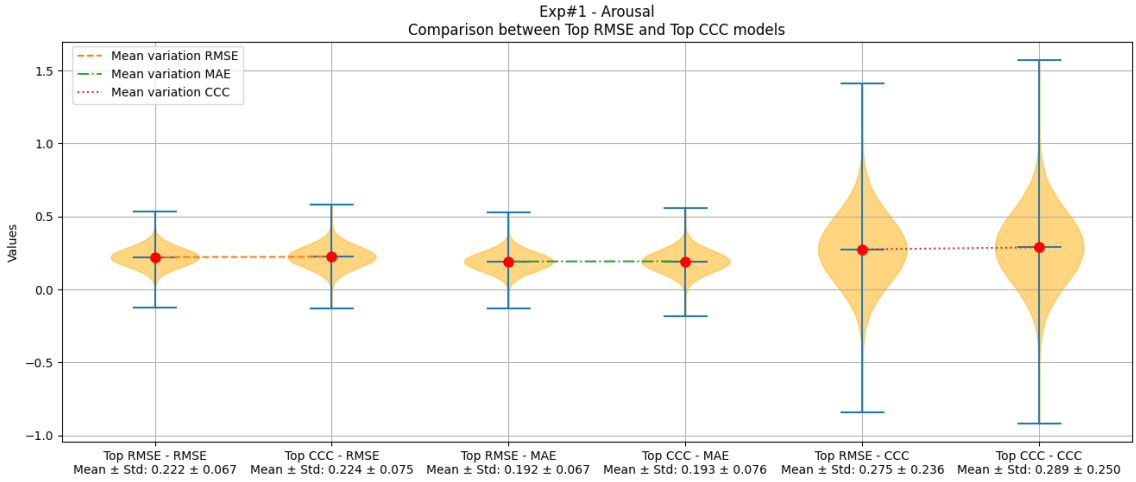


(b) Exp #2

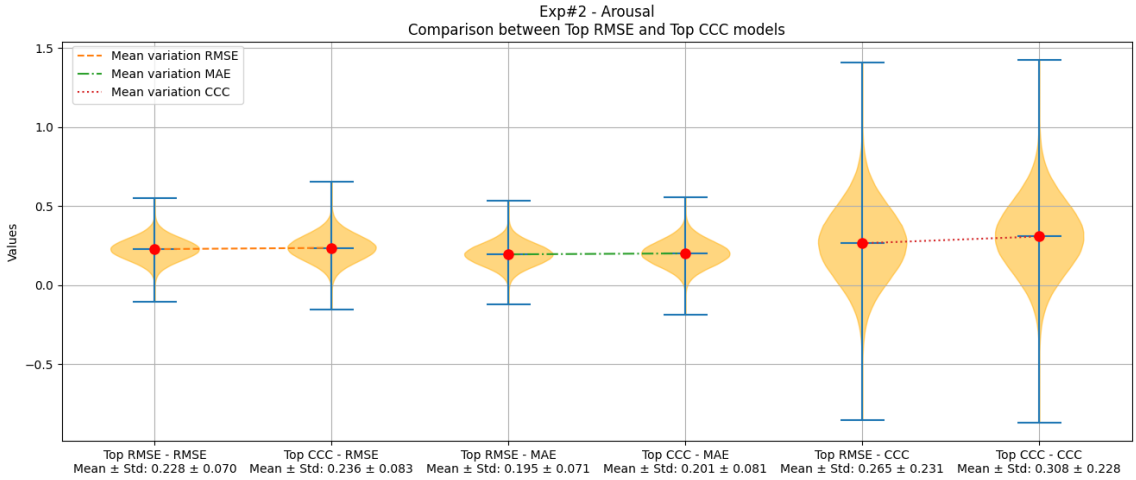


(c) Exp #3

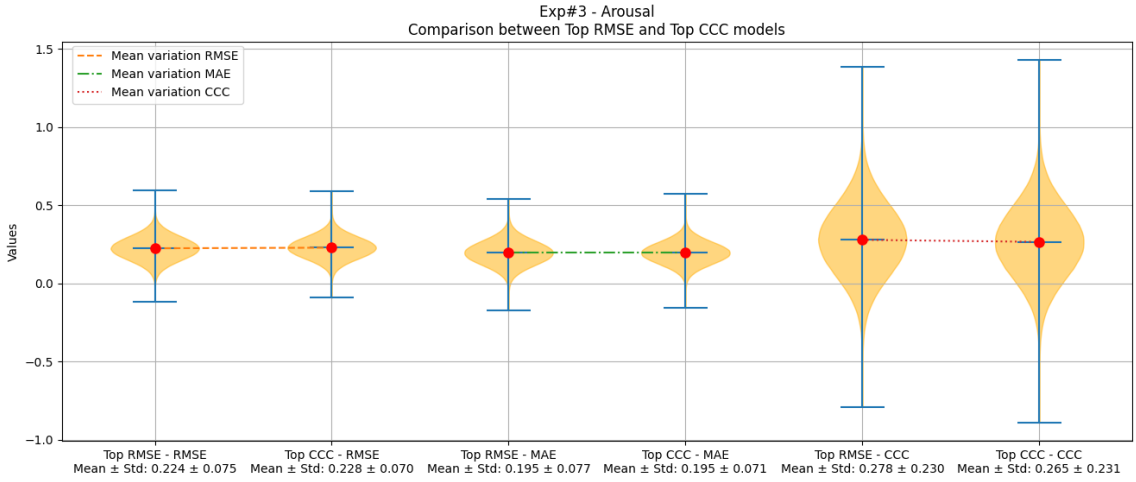
Figure 4.9: A Violin Plot comparison depicting the variation of RMSE, MAE, and CCC metrics for the Valence dimension based on top metric selection criteria.



(a) Exp #1



(b) Exp #2

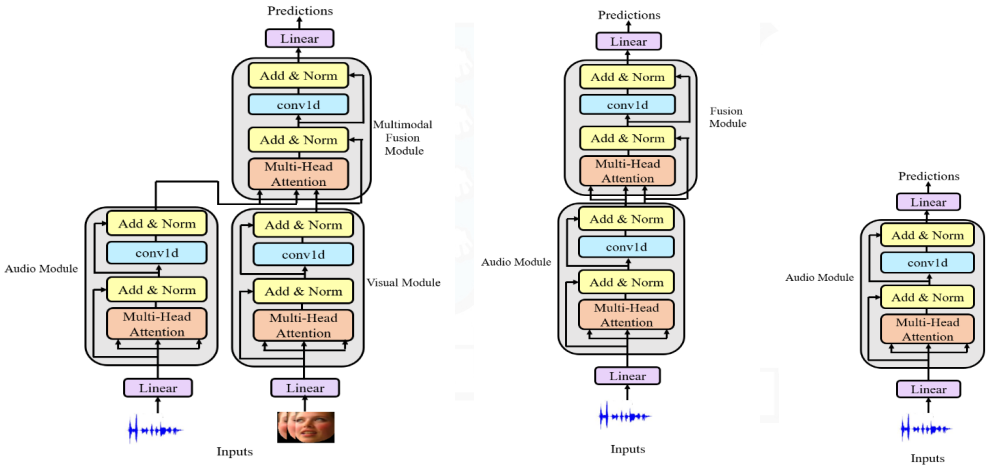


(c) Exp #3

Figure 4.10: A Violin Plot comparison depicting the variation of RMSE, MAE, and CCC metrics for the Arousal dimension based on top metric selection criteria.

4.3 Transformer Model Configurations

4.3.1 Model Definition



(a) Multimodal **(b)** Bimodular **(c)** Unimodular

Figure 4.11: In the left (4.11a), the multimodal Transformer model presented in by Huang *et al.* [14]. (4.11b, 4.11c) Transformer models proposed for MEVD analysis directly from music audio features by deconstructing the model.

In their 2022 work, Huang *et al.* [14] introduced a Multimodal Transformer-based Prediction System designed to fuse audio-visual modalities at different levels. The choice of a Transformer model was motivated by its ability to capture emotional long-term temporal dependencies through its self-attention mechanism. Their proposal focuses on achieving multimodal fusion at feature or decision levels, transforming high-level output representations of audio and visual modules into a common semantic feature space to produce effective multimodal feature representations (Fig 4.11a). The dataset used for this work, extracted from the AVEC dataset (2.4.4), underwent preprocessing with openSMILE, adhering to the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) settings, obtaining a set of 88-dimensional features. Annotations were available every 100$[ms]$, scaled between $[-1, 1]$, and underwent a PCA process with 95% variance retention, aligning with a preprocessing format similar to that performed for DEAM values. The evaluation measure selected in this research is CCC, which is presented for monomodal and multimodal configuration of their network architecture, thereby validating the viability of a deconstructive analysis at a modular level over this model.

Taking inspiration from this architecture, the visual and multimodal fusion modules were re-

moved, retaining only the audio module. The input layer was adjusted to fit the openSMILE-based DEAM feature set. Two main model architecture configurations were implemented, depending on whether to maintain an adapted fusion module to act as a second audio module layer. Additionally, in both cases, the inclusion of an LSTM layer before the final prediction layer was considered, based on the authors' suggestion that a model combining Transformer and LSTM models can better learn emotional temporal dependencies and achieve a promising increase in performance. The hypothesis here is that the *bimodular* approach (Fig. 4.11a) should yield better performance metrics than the *unimodular* approach (Fig. 4.11c) due to the presence of two continuous modules processing the values. Also, following the authors' statement, models with an LSTM layer should perform better.

A deconstructive analysis is presented by evaluating *unimodular* and *bimodular* configurations. The layered configuration of the model presented by Huang *et al.* and all reported network configuration parameters serve as the basis for the implemented models. The *unimodular* approach is based on a similar analysis performed by the authors, who also performed an unimodal continuous emotion recognition approach with both feature modalities to compare and validate their multimodal approach, arguing that unimodal models can achieve effective performance in this context. Multiple model parameter configurations for each modal approach are considered to find the most suited model configuration for each emotional dimension.

The nomenclature used to reference the results obtained from the implemented models, based on their architecture and model configuration, should be interpreted as follows: $\{A|V\}\{1|2\}\{L|N\}\{P|N\}\{A|S\}\_\{LR\}\_\{Epochs\}$ where

- **Module:** Unimodule or bimodule module structure. (1|2)
- **LSTM layer:** With or without final LSTM layer. (L|N)
- **PCA:** With or without PCA applied on the training set. (P|N)
- **Loss function:** MAE or MSE. (A|S)
- **Learning Rate:** {0.001, 0.010, 0.020, 0.030, 0.040, 0.050, 0.060, 0.070}.

- **Epochs:** 100, with model values captured at epochs {1,5,10,25,50,75,100}.

4.3.2 Implementation

In this Subsection, analogous to the analysis conducted for the MLP baseline, the configuration of the models with the best RMSE and CCC scores for each modular approach is presented. For deconstructive analysis purposes, the individual top scores for configurations with or without LSTM for each emotional dimension are indicated, along with plots displaying the estimations obtained over the evaluation dataframe for each top model and each score. Only models trained for at least 25 epochs were reported for statistical significance, while models having 1, 5, and 10 epochs were omitted. The implemented model architecture follows the design presented in [14] (Fig. 4.11a). All multi-head attention modules comprise two 4-head attention layers, each with a residual connection and layer normalization. The number of hidden nodes of the attention layer and the output channels of the conv1d layer are set to 64. Adam's optimization algorithm and a dropout layer with a 0.5 rate were used, consistent with Huang *et al.* implementation. The presented values correspond to the average of the metrics computed at a song level across songs.

Assuming that the expression of music emotion at any moment corresponds to the accumulation of the previous context before that moment in music, a rolling window average is applied over a padded copy of each song's predicted values. This smoothing technique aims to yield less noisy and more representative trajectory prediction values. The padding step repeats the first and last values as necessary, with the assumption that perceived emotion is more likely to retain its state than to start from a neutral state (zero padding) or even an inverse emotional state (reverse padding). Two cases were considered, with windows of size 5 (i.e., $\pm 1[s]$) and 10 (i.e., $\pm 2.5[s]$). The top RMSE models obtained exhibit consistent behavior across both window widths, and, as such, the window of width 10 timesteps is used and reported.

The code developed to implement the unimodular (Fig. 4.12) and bimodular (Fig. 4.14) architectures is presented in each section. The LSTM layer is included to provide a more comprehensive view of the code, and it is removed as needed during the analysis.

4.3.2.1 Unimodular

In this section, the top models obtained under RMSE (Table 4.14) and CCC (Table 4.15) criteria for the Unimodular approach are displayed. It is immediately noticeable that the CCC values for the arousal dimension are higher without significant loss in RMSE score. However, for the valence dimension, the increase in CCC values is much smaller compared to arousal values, accompanied by higher RMSE scores. Another noteworthy observation is the model configuration obtained for top models: in both emotional dimensions, models with better CCC scores have lower learning rate configuration values.

Figure 4.12 displays the code snippet of the implemented model architecture configuration, which depicts the MHA module immediately after a linear or dense layer. Figure 4.13 presents the obtained predictions for the top RMSE and CCC models in each LSTM configuration scenario.

Top RMSE

Table 4.14: Top 5 models with the lowest RMSE scores implemented in the Unimodular approach, comparing cases with and without the LSTM layer.

LSTM	Valence				Arousal			
	Model configuration	RMSE	MAE	CCC	Model configuration	RMSE	MAE	CCC
No	V1NPS_0.070_50	0.264±0.139	0.246±0.138	-0.000±0.000	A1NPA_0.030_50	0.233±0.080	0.202±0.080	0.192±0.183
	V1NPS_0.050_50	0.264±0.141	0.246±0.140	0.001±0.003	A1NPS_0.060_75	0.249±0.102	0.221±0.103	0.119±0.137
	V1NPA_0.070_25	0.264±0.135	0.247±0.134	-0.001±0.007	A1NPS_0.060_25	0.251±0.099	0.223±0.101	0.116±0.134
	V1NPS_0.060_75	0.264±0.132	0.247±0.131	0.001±0.002	A1NPS_0.050_25	0.252±0.116	0.221±0.116	0.150±0.169
	V1NPS_0.060_50	0.264±0.140	0.247±0.139	0.000±0.001	A1NNA_0.050_100	0.255±0.107	0.223±0.105	0.143±0.166
Yes	V1LPS_0.050_50	0.264±0.141	0.246±0.141	0.003±0.016	A1LPS_0.001_25	0.230±0.074	0.199±0.076	0.274±0.243
	V1LPS_0.060_75	0.264±0.141	0.246±0.141	0.002±0.009	A1LNA_0.001_75	0.234±0.085	0.201±0.087	0.311±0.251
	V1LPA_0.030_75	0.264±0.137	0.246±0.136	0.000±0.001	A1LPA_0.001_50	0.237±0.083	0.203±0.084	0.282±0.255
	V1LPA_0.040_25	0.264±0.135	0.246±0.134	0.000±0.000	A1LNS_0.010_75	0.238±0.090	0.206±0.092	0.317±0.253
	V1LPS_0.070_100	0.264±0.137	0.246±0.136	0.000±0.000	A1LNS_0.001_25	0.238±0.090	0.206±0.093	0.310±0.252

Top CCC

Table 4.15: Top 5 models with the highest CCC scores implemented in the Unimodular approach, comparing cases with and without the LSTM layer.

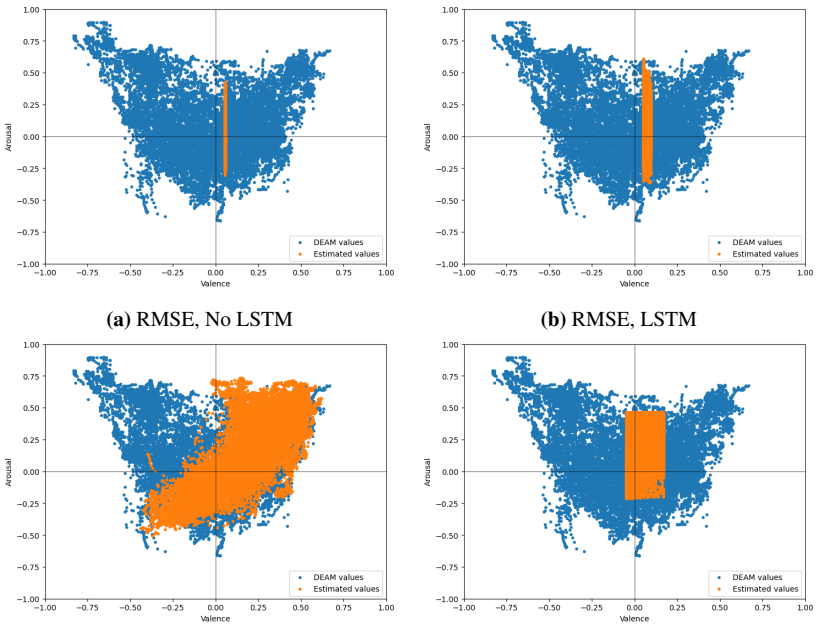
LSTM	Valence				Arousal			
	Model configuration	RMSE	MAE	CCC	Model configuration	RMSE	MAE	CCC
No	V1NPS_0.001_100	0.352±0.195	0.311±0.191	0.053±0.132	A1NNA_0.001_50	0.261±0.099	0.225±0.101	0.330±0.259
	V1NPS_0.001_25	0.341±0.191	0.303±0.186	0.052±0.136	A1NNS_0.001_100	0.277±0.100	0.238±0.103	0.316±0.248
	V1NPS_0.001_50	0.334±0.183	0.296±0.179	0.049±0.130	A1NPS_0.001_50	0.277±0.105	0.240±0.108	0.315±0.257
	V1NPA_0.001_25	0.340±0.184	0.303±0.181	0.048±0.139	A1NNS_0.001_25	0.260±0.099	0.225±0.102	0.315±0.258
	V1NPA_0.020_100	0.316±0.194	0.285±0.190	0.048±0.144	A1NPA_0.001_50	0.274±0.104	0.238±0.107	0.314±0.257
Yes	V1LPS_0.001_50	0.292±0.155	0.256±0.149	0.064±0.135	A1LNS_0.010_75	0.238±0.090	0.206±0.092	0.317±0.253
	V1LPS_0.001_100	0.303±0.158	0.266±0.153	0.062±0.143	A1LNA_0.001_50	0.238±0.082	0.203±0.085	0.312±0.238
	V1LNS_0.001_75	0.285±0.169	0.256±0.165	0.061±0.131	A1LNS_0.001_50	0.238±0.089	0.206±0.092	0.312±0.250
	V1LPS_0.001_75	0.303±0.157	0.265±0.152	0.060±0.138	A1LNA_0.001_75	0.234±0.085	0.201±0.087	0.311±0.251
	V1LNS_0.001_50	0.292±0.161	0.256±0.155	0.059±0.128	A1LNS_0.001_25	0.238±0.090	0.206±0.093	0.310±0.252

Code Snippet

```
Layer (type)                Output Shape              Param #   Connected to
==================================================================================================
input_1 (InputLayer)        [(None, 84)]              0         []

dense (Dense)               (None, 64)                5440      ['input_1[0][0]']

multi_head_attention (Mult  ((None, None, 64),        16640     ['dense[0][0]',
iHeadAttention)              (None, 4, None, None))               'dense[0][0]',
                                                                  'dense[0][0]']

dropout (Dropout)           (None, None, 64)          0         ['multi_head_attention[0][0]']

add (Add)                   (None, None, 64)          0         ['dense[0][0]',
                                                                  'dropout[0][0]']

layer_normalization (Layer  (None, None, 64)          128       ['add[0][0]']
Normalization)

dense_5 (Dense)             (None, None, 64)          4160      ['layer_normalization[0][0]']

dropout_1 (Dropout)         (None, None, 64)          0         ['dense_5[0][0]']

add_1 (Add)                 (None, None, 64)          0         ['layer_normalization[0][0]',
                                                                  'dropout_1[0][0]']

layer_normalization_1 (Lay  (None, None, 64)          128       ['add_1[0][0]']
erNormalization)

lstm (LSTM)                 (None, 64)                33024     ['layer_normalization_1[0][0]']

dense_6 (Dense)             (None, 1)                 65        ['lstm[0][0]']
==================================================================================================
Total params: 59585 (232.75 KB)
Trainable params: 59585 (232.75 KB)
Non-trainable params: 0 (0.00 Byte)
```

Figure 4.12: Unimodular model architecture with LSTM.

Top models



(a) RMSE, No LSTM **(b)** RMSE, LSTM

(c) CCC, No LSTM **(d)** CCC, LSTM

Figure 4.13: Visualization of the predictions obtained with the unimodular approach for top RMSE and CCC models.

4.3.2.2 Bimodular

In this section, the top models obtained under RMSE (Table 4.16) and CCC (Table 4.17) criteria for the bimodular approach are displayed. Similar to the unimodal approach, it is immediately noticeable that the CCC values for the arousal dimension are higher without significant loss in RMSE score. However, for the valence dimension, the increase in CCC values is much smaller compared to arousal values, accompanied by higher RMSE scores. Another noteworthy observation is the model configuration obtained for top models: in both emotional dimensions, models with better CCC scores have lower learning rate configuration values. Figure 4.14 displays the code snippet of the implemented model architecture configuration, which depicts the first MHA module immediately after a linear or dense layer, with the second module stacked after the layer normalization step. Figure 4.13 presents the obtained predictions for the top RMSE and CCC models in each LSTM configuration scenario.

Top RMSE

Table 4.16: Top 5 models with the lowest RMSE scores implemented in the Bimodular approach, comparing cases with and without the LSTM layer.

LSTM	Valence				Arousal			
	Model configuration	RMSE	MAE	CCC	Model configuration	RMSE	MAE	CCC
No	V2NPS_0.040_75	0.264±0.133	0.247±0.131	0.004±0.010	A2NPS_0.040_25	0.247±0.104	0.221±0.104	0.071±0.092
	V2NPS_0.060_100	0.264±0.133	0.246±0.132	0.000±0.000	A2NNS_0.030_75	0.251±0.112	0.222±0.110	0.120±0.151
	V2NPS_0.070_50	0.264±0.133	0.246±0.132	-0.000±0.000	A2NNA_0.001_25	0.253±0.095	0.218±0.097	0.323±0.261
	V2NNA_0.070_25	0.264±0.135	0.246±0.134	0.004±0.017	A2NNS_0.001_25	0.253±0.091	0.219±0.093	0.314±0.249
	V2NNS_0.030_50	0.264±0.137	0.246±0.136	0.004±0.018	A2NNA_0.001_100	0.255±0.099	0.220±0.102	0.351±0.261
Yes	V2LNS_0.050_25	0.264±0.136	0.246±0.135	0.000±0.000	A2LPS_0.001_50	0.223±0.075	0.194±0.077	0.286±0.237
	V2LPS_0.050_75	0.264±0.135	0.246±0.134	0.000±0.000	A2LPS_0.001_75	0.225±0.077	0.196±0.079	0.281±0.232
	V2LNS_0.030_75	0.264±0.137	0.246±0.136	0.000±0.000	A2LPS_0.001_25	0.226±0.081	0.198±0.083	0.291±0.231
	V2LNA_0.020_100	0.264±0.134	0.246±0.133	0.000±0.000	A2LPS_0.001_100	0.227±0.079	0.197±0.081	0.287±0.233
	V2LNS_0.020_25	0.264±0.133	0.246±0.132	0.000±0.000	A2LNS_0.001_25	0.228±0.085	0.200±0.088	0.311±0.244

Top CCC

Table 4.17: Top 5 models with the highest CCC scores implemented in the Bimodular approach, comparing cases with and without the LSTM layer.

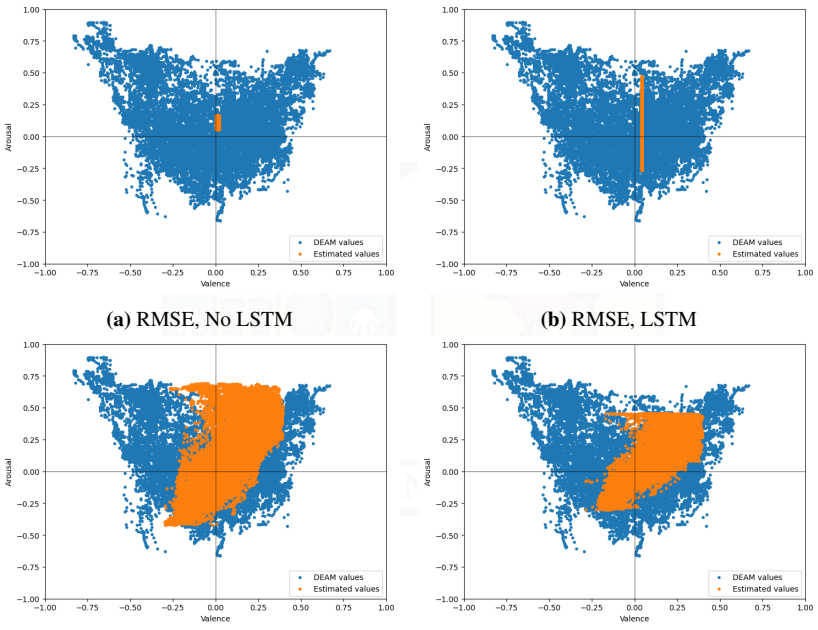
LSTM	Valence				Arousal			
	Model configuration	RMSE	MAE	CCC	Model configuration	RMSE	MAE	CCC
No	V2NNS_0.001_25	0.297±0.165	0.263±0.162	0.053±0.144	A2NNA_0.001_100	0.255±0.099	0.220±0.102	0.351±0.261
	V2NPS_0.001_75	0.333±0.190	0.296±0.186	0.051±0.139	A2NNA_0.001_25	0.253±0.095	0.218±0.097	0.323±0.261
	V2NNS_0.001_50	0.322±0.175	0.284±0.171	0.049±0.136	A2NPA_0.001_50	0.276±0.106	0.238±0.110	0.318±0.255
	V2NPS_0.001_50	0.339±0.197	0.302±0.193	0.049±0.136	A2NNS_0.001_100	0.262±0.086	0.223±0.090	0.316±0.254
	V2NPS_0.001_25	0.329±0.189	0.293±0.185	0.049±0.146	A2NNS_0.001_25	0.253±0.091	0.219±0.093	0.314±0.249
Yes	V2LNS_0.001_50	0.290±0.169	0.258±0.165	0.060±0.141	A2LNS_0.001_50	0.235±0.091	0.206±0.093	0.317±0.252
	V2LNS_0.001_75	0.289±0.164	0.257±0.161	0.057±0.134	A2LNS_0.001_25	0.228±0.085	0.200±0.088	0.311±0.244
	V2LPS_0.001_25	0.296±0.176	0.267±0.173	0.055±0.146	A2LNS_0.001_75	0.228±0.082	0.200±0.085	0.305±0.251
	V2LPS_0.001_100	0.295±0.172	0.265±0.169	0.053±0.138	A2LNS_0.001_100	0.232±0.082	0.204±0.085	0.297±0.251
	V2LNS_0.001_25	0.294±0.167	0.260±0.163	0.052±0.140	A2LPA_0.001_75	0.244±0.093	0.214±0.096	0.295±0.249

Code Snippet

```
Layer (type)                Output Shape              Param #   Connected to
==================================================================================================
input_1 (InputLayer)        [(None, 84)]              0         []

dense (Dense)               (None, 64)                5440      ['input_1[0][0]']

multi_head_attention (Mult  ((None, None, 64),        16640     ['dense[0][0]',
iHeadAttention)              (None, 4, None, None))               'dense[0][0]',
                                                                  'dense[0][0]']

dropout (Dropout)           (None, None, 64)          0         ['multi_head_attention[0][0]']

add (Add)                   (None, None, 64)          0         ['dense[0][0]','dropout[0][0]']

layer_normalization (Layer  (None, None, 64)          128       ['add[0][0]']
Normalization)

dense_5 (Dense)             (None, None, 64)          4160      ['layer_normalization[0][0]']

dropout_1 (Dropout)         (None, None, 64)          0         ['dense_5[0][0]']

add_1 (Add)                 (None, None, 64)          0         ['layer_normalization[0][0]',
                                                                  'dropout_1[0][0]']

layer_normalization_1 (Lay  (None, None, 64)          128       ['add_1[0][0]']
erNormalization)

multi_head_attention_1 (Mu  ((None, None, 64),        16640     ['layer_normalization_1[0][0]',
ltiHeadAttention)            (None, 4, None, None))               'layer_normalization_1[0][0]',
                                                                  'layer_normalization_1[0][0]'
                                                                 ]

dropout_2 (Dropout)         (None, None, 64)          0         ['multi_head_attention_1[0][0]']

add_2 (Add)                 (None, None, 64)          0         ['layer_normalization_1[0][0]',
                                                                  'dropout_2[0][0]']

layer_normalization_2 (Lay  (None, None, 64)          128       ['add_2[0][0]']
erNormalization)

dense_10 (Dense)            (None, None, 64)          4160      ['layer_normalization_2[0][0]']

dropout_3 (Dropout)         (None, None, 64)          0         ['dense_10[0][0]']

add_3 (Add)                 (None, None, 64)          0         ['layer_normalization_2[0][0]',
                                                                  'dropout_3[0][0]']

layer_normalization_3 (Lay  (None, None, 64)          128       ['add_3[0][0]']
erNormalization)

lstm (LSTM)                 (None, 64)                33024     ['layer_normalization_3[0][0]']

dense_11 (Dense)            (None, 1)                 65        ['lstm[0][0]']
==================================================================================================
Total params: 80641 (315.00 KB)
Trainable params: 80641 (315.00 KB)
Non-trainable params: 0 (0.00 Byte)
```

Figure 4.14: Bimodular model architecture with LSTM.

Top models



(a) RMSE, No LSTM **(b)** RMSE, LSTM

(c) CCC, No LSTM **(d)** CCC, LSTM

Figure 4.15: Visualization of the predictions obtained with the Bimodular approach for top RMSE and CCC models.

4.3.3 Model Analysis

The Figures 4.12 (Unimodular) and 4.14 (Bimodular) emphasize that relying solely on RMSE scores may not ensure optimal model prediction performance, particularly in the context of MEVD approaches. Conversely, the CCC metric demonstrates promising results. For instance, the prediction plots for the Unimodal No-LSTM model with highest CCC score (Fig. 4.13c) align with the shape of the dataframe used for training. However, this alignment could potentially indicate overfitting or prediction biases due to imbalances in the training set.

The models displayed in Table 4.18 correspond to the best model of each configuration combination tested. Following the analysis performed in the previous section, Figures 4.16 (Unimodular) and 4.17 (Bimodular) depict the variation in RMSE, MAE and CCC metrics between the top models and their counterparts for the modular and optional LSTM architecture configurations in both emotional dimensions.

The obtained model's metrics outperforms the values from the MLP Regressor, with the bimodular model achieving the lowest RMSE scores: 0.223 ± 0.075 for Arousal with

an LSTM layer and 0.264 ± 0.133 for the Valence dimension without the LSTM layer, both using the dataset that underwent PCA. However, the models chosen to develop the MRS were the ones with highest CCC score, which yield 0.225 ± 0.099 for Arousal and 0.292 ± 0.135 for Valence dimensions.

Table 4.18: Top models with the best scores implemented in Unimodular and Bimodular approaches.

Metric	Valence				Arousal			
	Model configuration	RMSE	MAE	CCC	Model configuration	RMSE	MAE	CCC
RMSE	V1NPS_0.070_50	0.264±0.139	0.246±0.138	-0.000±0.000	A1NPA_0.030_50	0.233±0.080	0.202±0.080	0.192±0.183
	V1LPS_0.050_50	0.264±0.141	0.246±0.141	0.003±0.141	A1LPS_0.001_25	0.230±0.074	0.199±0.076	0.274±0.243
	V2NPS_0.040_75	0.264±0.133	0.247±0.131	0.004±0.010	A2NPS_0.040_25	0.247±0.104	0.221±0.104	0.071±0.092
	V2LNS_0.050_25	0.264±0.136	0.246±0.135	0.000±0.000	A2LPS_0.001_50	0.223±0.075	0.194±0.077	0.286±0.237
CCC	V1NPS_0.001_100	0.352±0.195	0.311±0.191	0.053±0.132	A1NNA_0.001_50	0.261±0.099	0.225±0.101	0.330±0.259
	V1LPS_0.001_50	0.292±0.155	0.256±0.149	0.064±0.135	A1LNS_0.010_75	0.238±0.090	0.206±0.092	0.317±0.253
	V2NNS_0.001_25	0.297±0.165	0.263±0.162	0.053±0.144	A2NNA_0.001_100	0.255±0.099	0.220±0.102	0.351±0.261
	V2LNS_0.001_50	0.290±0.169	0.258±0.165	0.060±0.141	A2LNS_0.001_50	0.235±0.091	0.206±0.093	0.295±0.249

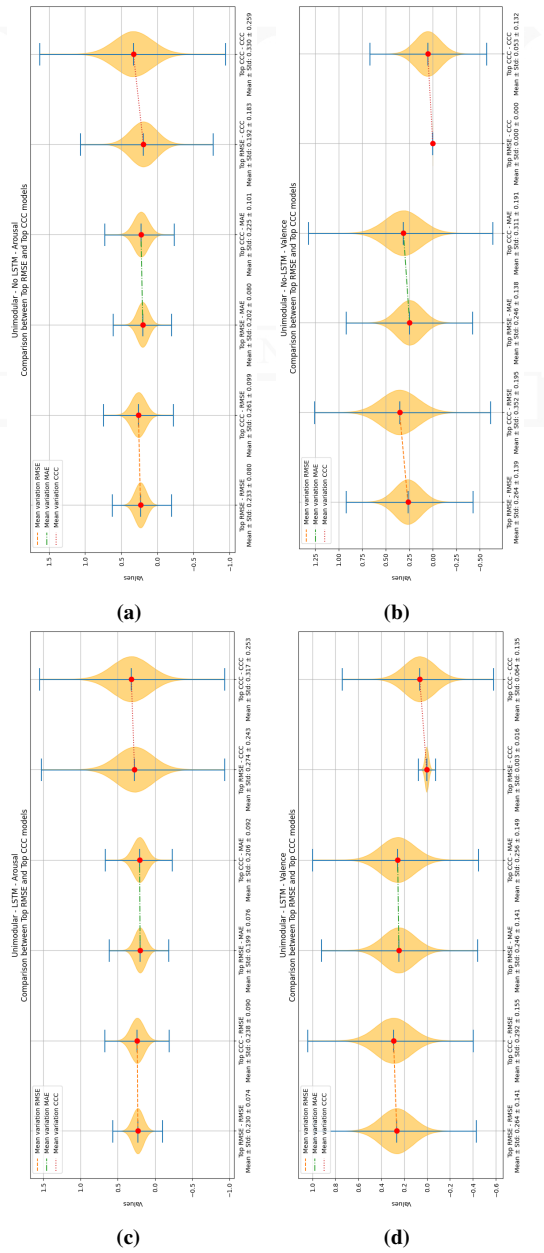


Figure 4.16: A Violin Plot comparison depicting the variation of RMSE, MAE, and CCC metrics for the Unimodular approach dimension based on top metric selection criteria.

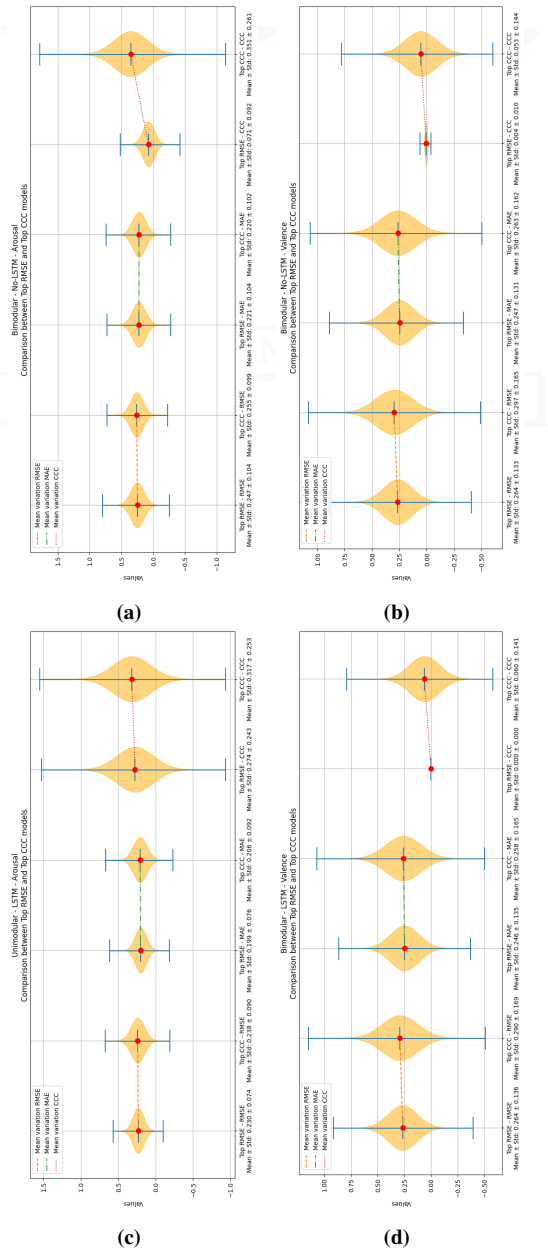


Figure 4.17: A Violin Plot comparison depicting the variation of RMSE, MAE, and CCC metrics for the Bimodular approach dimension based on top metric selection criteria.

5 | Emotional System

This chapter employs the emotional prediction values from the top Transformer models selected in the previous chapter as input for designing the Recommender System. Additionally, values from the MLP baseline models and the reference DEAM prediction values are incorporated for comparison and deeper analysis. The aim is to explore further the relationship between the CCC metric and the emotional prediction values of songs, represented as the *Emotional Trajectory* values. This concept, detailed in Section 2.3, entails visualizing predicted Valence/Arousal values over time for a specific song. In the ideal scenario, the goal is to reconstruct the emotional curve as accurately as possible. Then, by lowering the expected estimation resolution values, it becomes possible to use data obtained from the estimation of the emotional behavior of a song. This process includes validating the emotional changes over time under a discretization approach that reduces the emotional dimensional variety to a set of different classes or categories, encompassing certain emotional intervals over the plane.

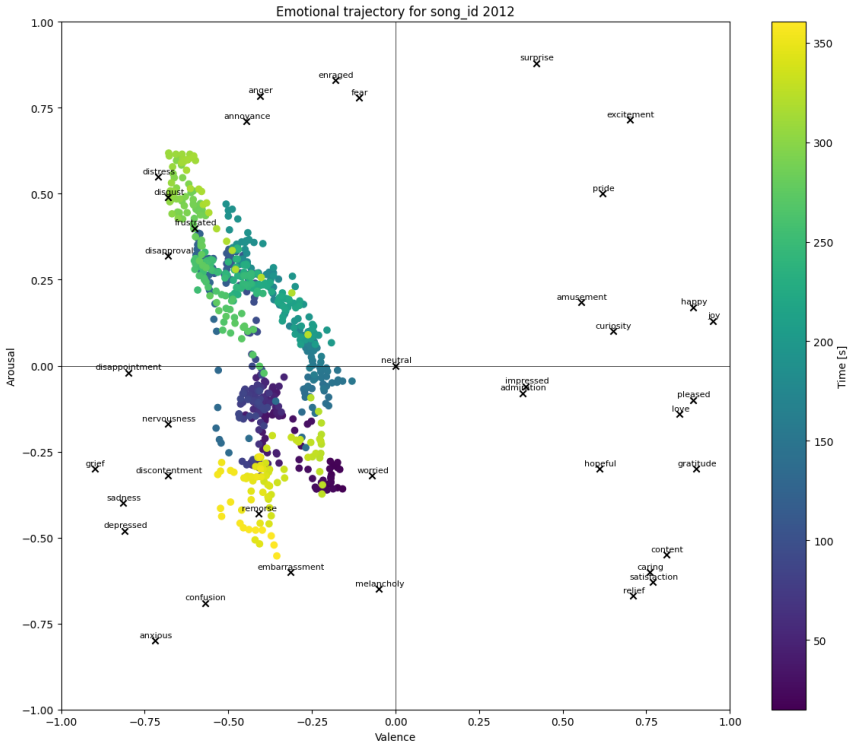
The Emotional Trajectory examination serves as a means to validate the prediction behavior of the models at a song-level specificity. While the reported metrics facilitated the assessment of the average prediction behavior of the models compared to state-of-the-art models, such comparisons lack the granularity offered by the emotional trajectory. This trajectory allows for identifying transitions, enabling a more nuanced characterization of a particular song's behavior. To delve deeper into this analysis, a method for evaluating emotional trajectory behavior is proposed, involving the application of a windowed version of the CCC metric. The underlying assumption is that this approach provides more direct information about the model's reported reconstruction of emotional trajectory values and

their change intensity.

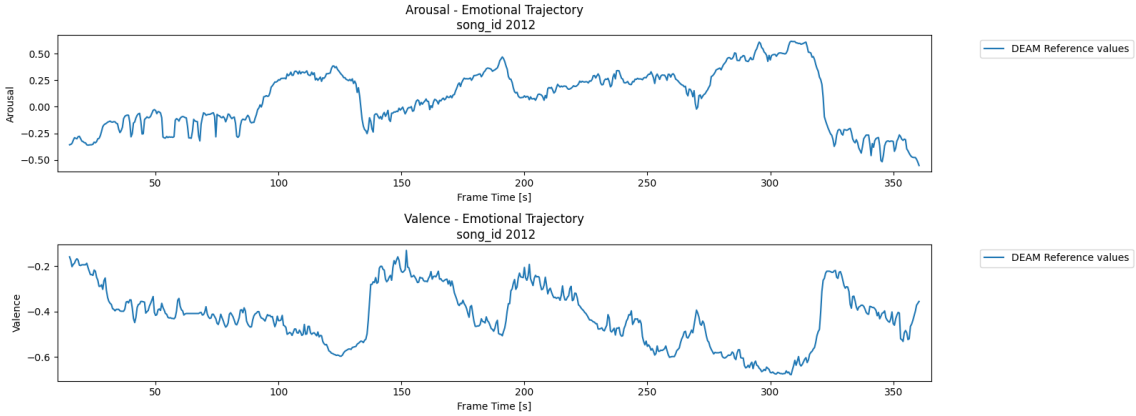
5.1 Emotional Trajectory Estimation

Considering that the Emotional Trajectory serves as a temporal visualization of predicted dimensional values on the V/A plane, the benefits of separately examining these dimensions lie in the detailed analysis of each dimension's prediction behavior. This approach is particularly relevant as independent models were trained for each dimension over different sets of objective values. Individually analyzing the dimensions allows for a comprehensive evaluation of the model behavior specific to each dimension. Furthermore, this approach enables the visualization of the expected behavior of the reference values. For instance, Figures 5.1 (Song # 2012) and 5.2 (Song # 2027) depict the reference Emotional Trajectory values available in DEAM dataset. These reference values represent the expected Emotional Trajectory over which the CCC metrics are computed. Subsequently, Figures 5.3 (Song # 2012) and 5.4 (Song # 2027) presents the V/A plot resulting from the estimations obtained with the Transformer models that achieved the highest CCC score for each dimension. Additionally, these figures offer a comparative dimensional behavior of the top metrics models for each approach.

5.1.1 Reference values

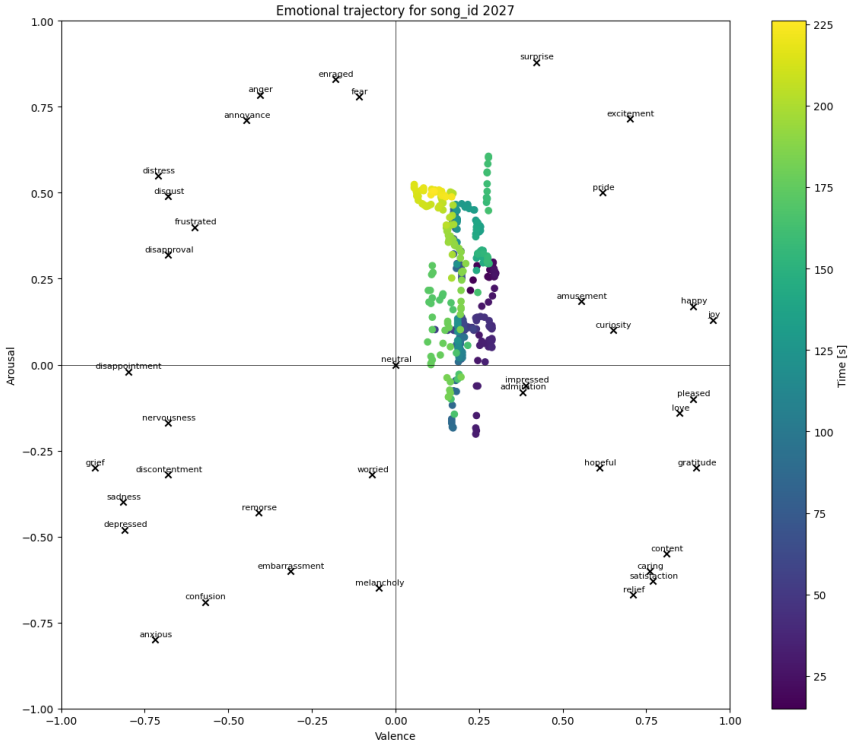


(a) V/A Emotional Trajectory Plot

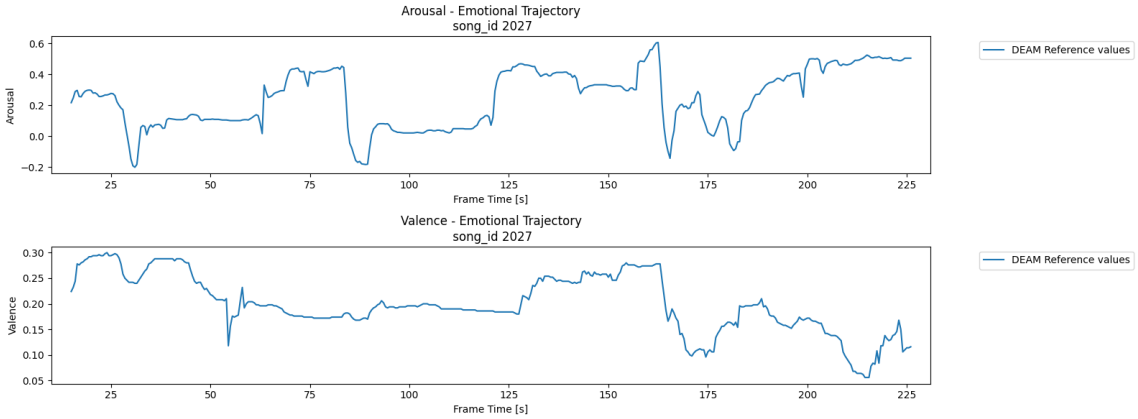


(b) Emotional Trajectory Plot by Dimension

Figure 5.1: Referential Emotional Trajectory Visualization for Song # 2012.



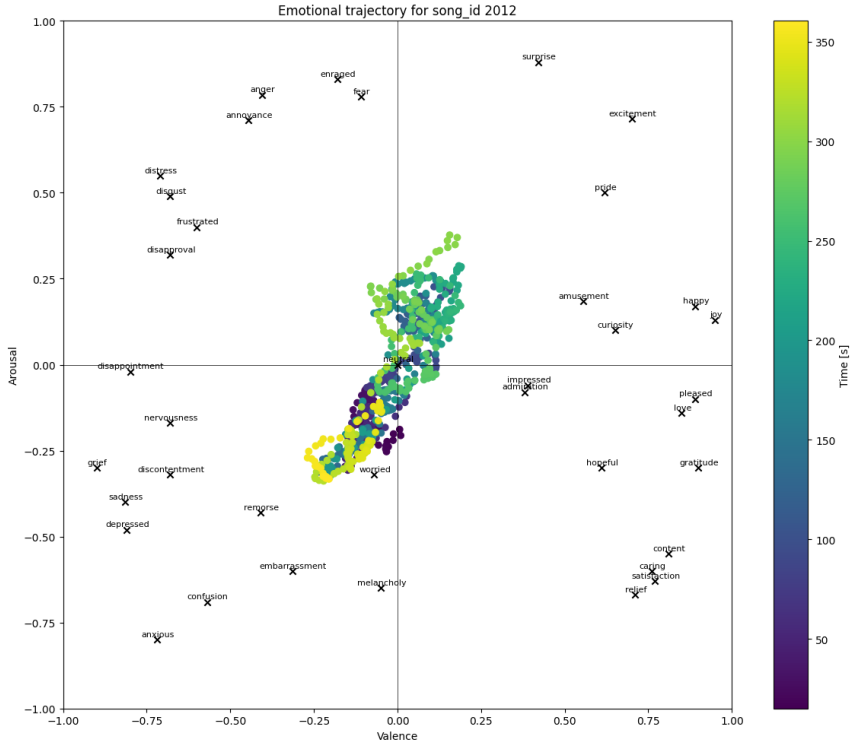
(a) V/A Emotional Trajectory Plot



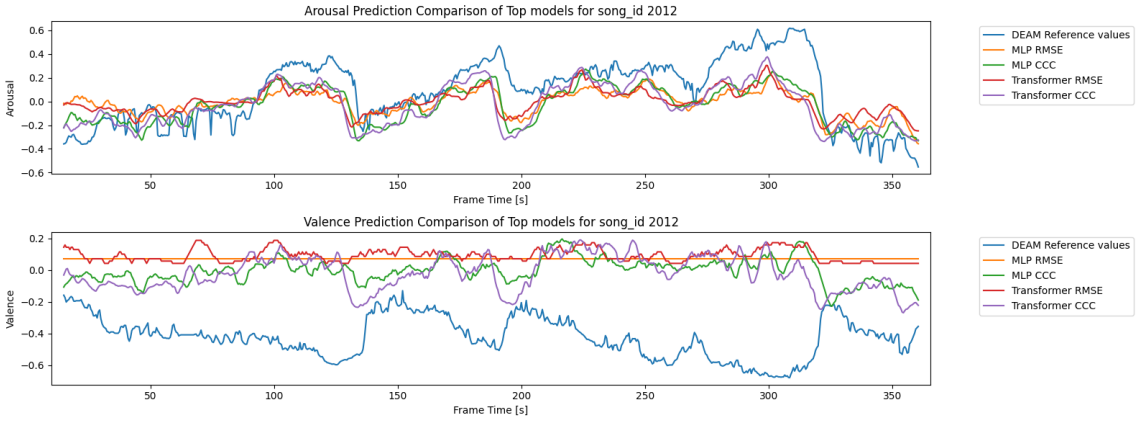
(b) Emotional Trajectory Plot by Dimension

Figure 5.2: Referential Emotional Trajectory Visualization for Song # 2027.

5.1.2 Estimated values

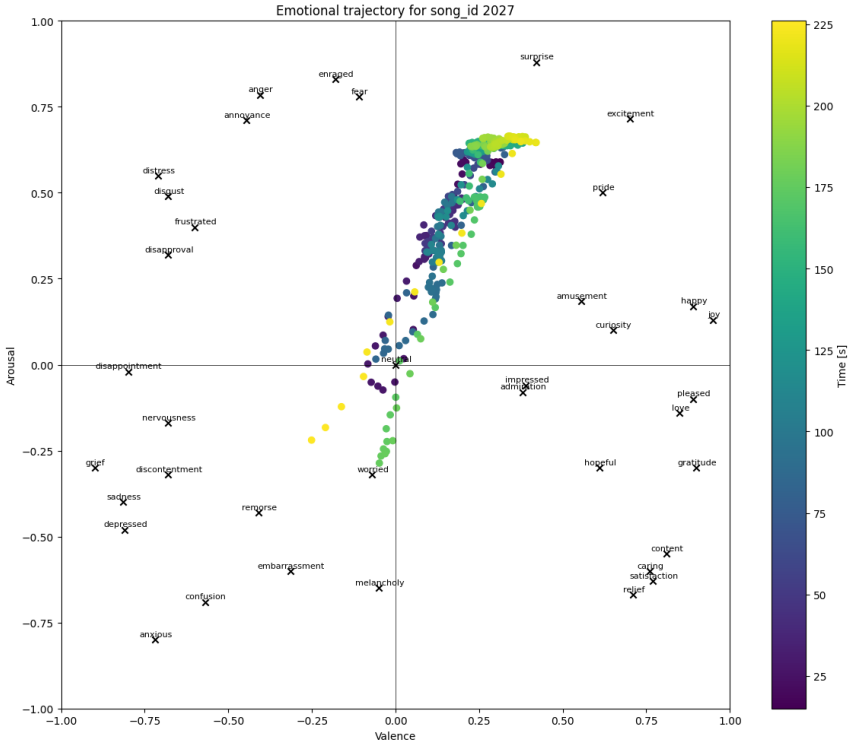


(a) V/A Estimated Emotional Trajectory Plot

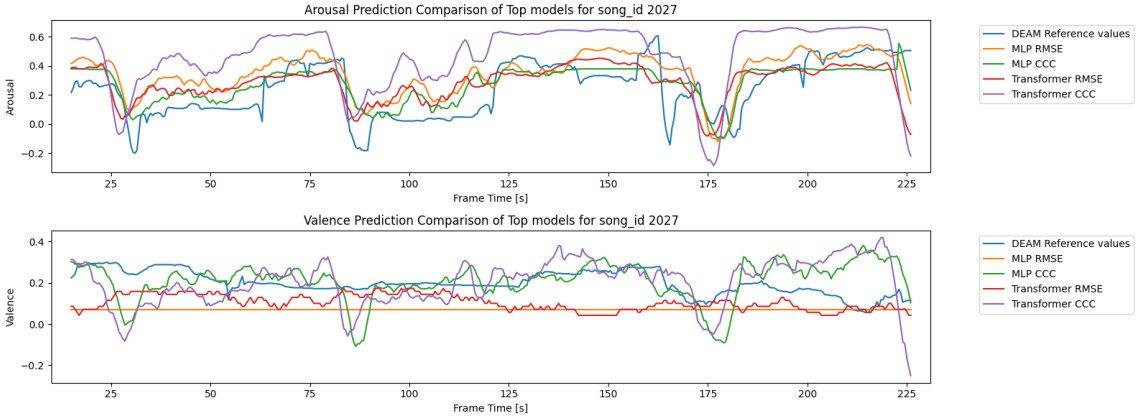


(b) Estimated Emotional Trajectory Plot by Dimension

Figure 5.3: Estimated Emotional Trajectory Visualization for Song # 2012.



(a) V/A Estimated Emotional Trajectory Plot



(b) Estimated Emotional Trajectory Plot by Dimension

Figure 5.4: Estimated Emotional Trajectory Visualization for Song # 2027.

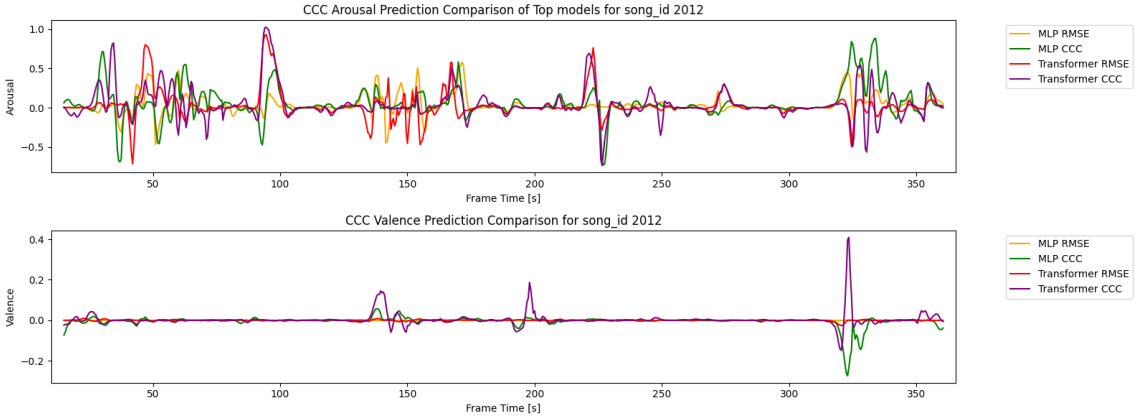
5.1.3 Estimation Analysis

After scrutinizing the prediction comparisons for all songs in the evaluation set, a notable pattern emerged: the model predictions tended to align with the shape of the Emotional Trajectory signal, albeit with a shift in their prediction values. Notably, in the valence dimension predictions, the values often exhibited a mirrored behavior in some instances, preserving the prediction of transitions within the trajectory. To analyze this phenomenon, a windowed CCC metric with a sliding window of size 10, aligning with the same assumption employed in the postprocessing step, was applied. This metric plays a crucial role in identifying correlations and discrepancies between the predicted and reference emotional trajectories, thereby enhancing the understanding of the model's prediction behavior. This implies that the predicted trajectory behavior replicates the shape and is well-positioned in the corresponding emotional dimension.

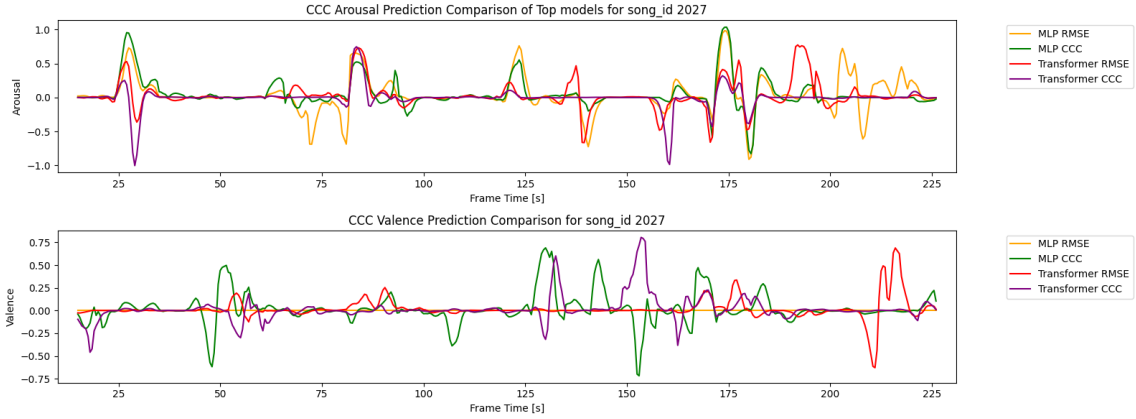
This metric allows to visualize both correlation and the direction or gradient of the compared emotional signals. For instance, if the Emotional Trajectory reference value in a specific dimension displays a positive incremental behavior, aligning with the predicted values behavior, the windowed metric shows a value close to 1. This indicates that not only do the predicted values exhibit similar incremental behavior but also that the values are well-positioned in time. Conversely, a windowed score close to -1 corresponds to the reverse scenario. In the valence dimension, this metric facilitates the visualization of instances when both signals exhibit similar behavior, regardless of the low accuracy presented by the models. For example, the windowed CCC metric for Song # 2012 predictions (Fig. 5.5a) displays for the arousal predictions two peaks at $100[s]$ and $225[s]$. The first peak matches with both the reference and predicted top CCC Transformer model, having very similar values and direction variation. Then, for the second timestamp mentioned it is possible to visualize two consecutive peaks with opposite sign, which are consistent with the matching trajectories in that time interval on which the Transformer CCC prediction changed its direction. A similar but smaller peak, due to lower interceptions between the signals, can be seen near the $250[s]$ mark. This is easier to discern for Song # 2027 (Fig. 5.5b) as the Emotional Trajectory presents smoother values for the arousal dimension. Here, the visible peaks in the windowed CCC metric coincide with the interception points,

where the negative peaks (e.g., $27[s]$, $165[s]$, $172[s]$) indicate an opposite gradient for the Emotional Trajectory values.

The observed behavior of the Emotional Trajectory estimation values showed that transitions could be reconstructed or identified to some extent, as the predicted signal reflects the expected reference emotional behavior. However, there is still considerable room for improvement in this context. For instance, the information derived from the proposed windowed CCC analysis could be leveraged as a new criterion in the training steps of different models, serving as a customized loss metric or weight factor. This approach is reminiscent of the work of Weniger *et al.* [64]. Integrating this metric into the training process could potentially lead to signals with higher correlations, thereby ensuring better prediction capabilities within this context. The primary advantage of employing this metric lies in its ability to provide a song-level correlation analysis in temporal emotional behavior, making it a valuable consideration for refining and enhancing model predictions.



(a) Song # 2012



(b) Song # 2027

Figure 5.5: Windowed CCC metric values visualization for Songs # 2012 and 2027.

5.2 Recommender System

The primary objective of a Music Recommender System Application is to offer users songs or playlists that align with specific input conditions, such as song titles, genres, or a particular song for recommendations based on similarity. In the context of Emotional Music Recommender Systems, users can input a specific emotion or a set of emotions to explore songs tagged under those categories. However, for a more dynamic approach, users should have the option to input a specific trajectory of emotions they wish to perceive expressed in the song or in a fragment of it. Another way to express user input is through the behavior of the emotional trajectory and the changes in the shape of the signal, by allowing the user to draw or select the desired trajectory values. This prediction can be utilized to represent the behavior of a particular piece of music. Subsequently, the system could extract representative segments of this trajectory, similar to the 45[s] long song excerpts available in the DEAM dataset. This dynamic input approach enables users to specify not only the target emotions but also the desired variation, or even patterns, of those emotions over time, leading to a more nuanced and personalized music recommendation experience.

Depending on the desired level of specificity associated with user input and the corresponding emotional values, these dimensional values can be discretized to facilitate the analysis of the desired Emotional Trajectory. Emotional values can be categorized by dividing the emotional plane into segments, similar to the four quadrants used in MER literature (HAHV, HALV, LAHV, LALV) (Fig. 5.6). This categorization is extended by adding the "Medium" segment, including Medium-High and Medium-Low Arousal/Valence values. Figures 5.7 (Song # 2012) and 5.8 (Song # 2027) shows the discretization of the referential emotional values by dimension, while Figures 5.9 (Song # 2012) and 5.10 (Song # 2027) displays the discretized predicted values for each song. Including "Medium" category is recommended to diminish the losing of valuable information during this discretization step, however it also adds more dimensional combinations to consider when estimating the trajectory statistics. Subsequently, the time spent over each categorical segment and the full song distribution of these values, along with the sequence describing this behavior, become

available as indicators. These indicators can be exploited as new category labels to generate recommendations based on the estimation of the Emotional Trajectory as input.

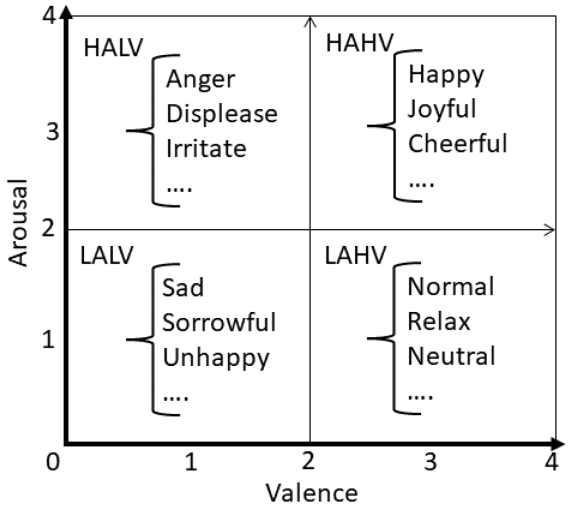
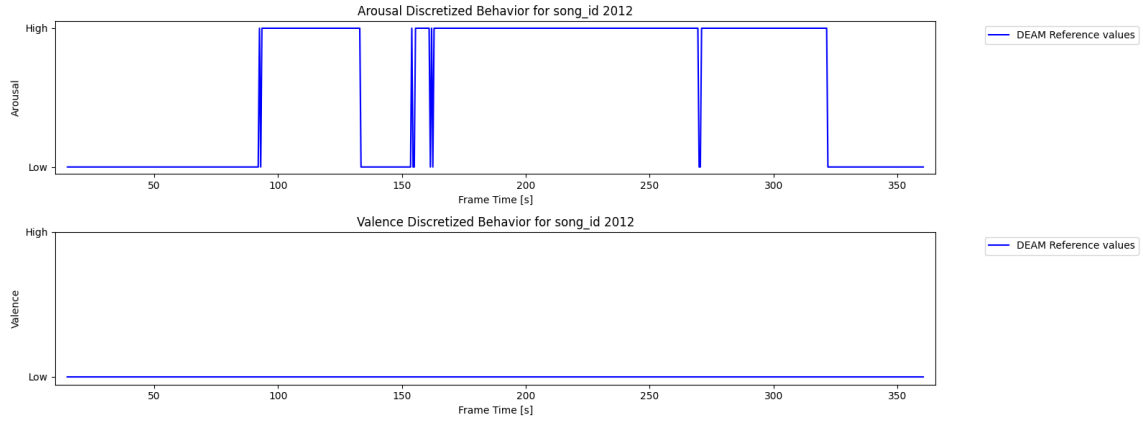
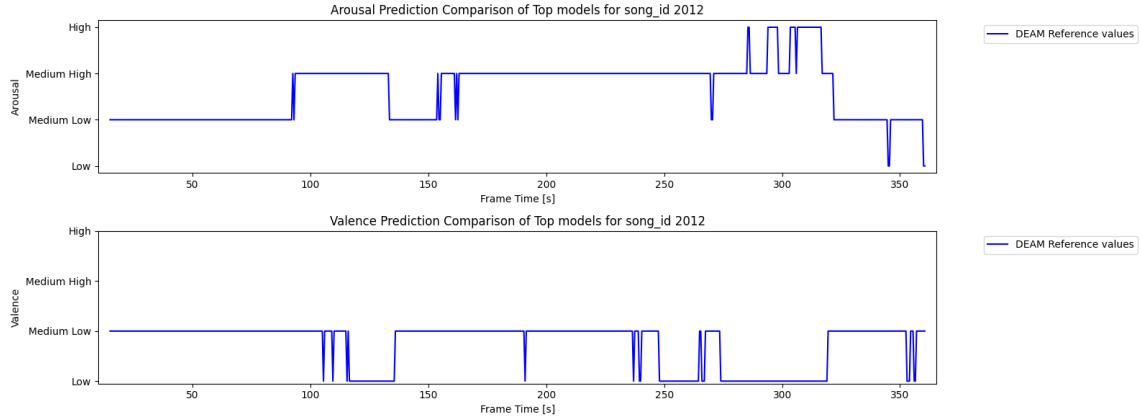


Figure 5.6: Emotion distribution and class distribution of the emotional tags based on the Russell circumplex model. Figure drawn from [15].

Reference

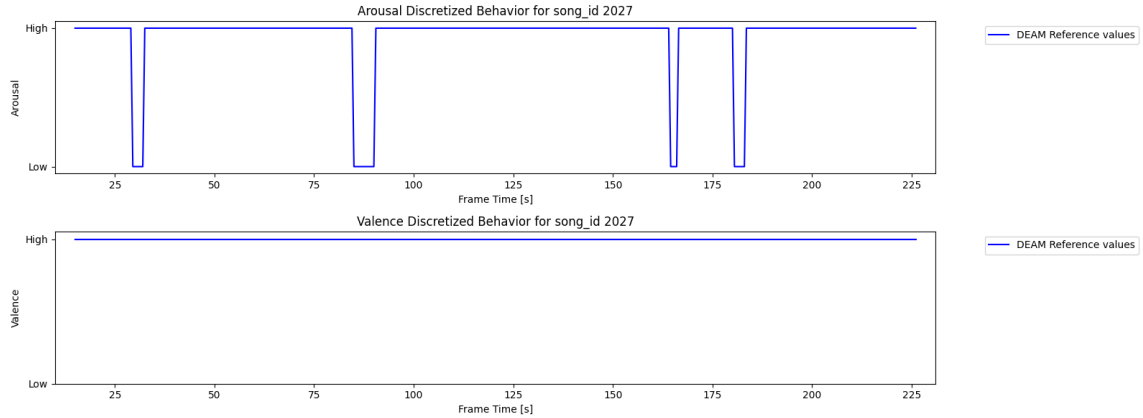


(a) Reference values discretized to High and Low categories.

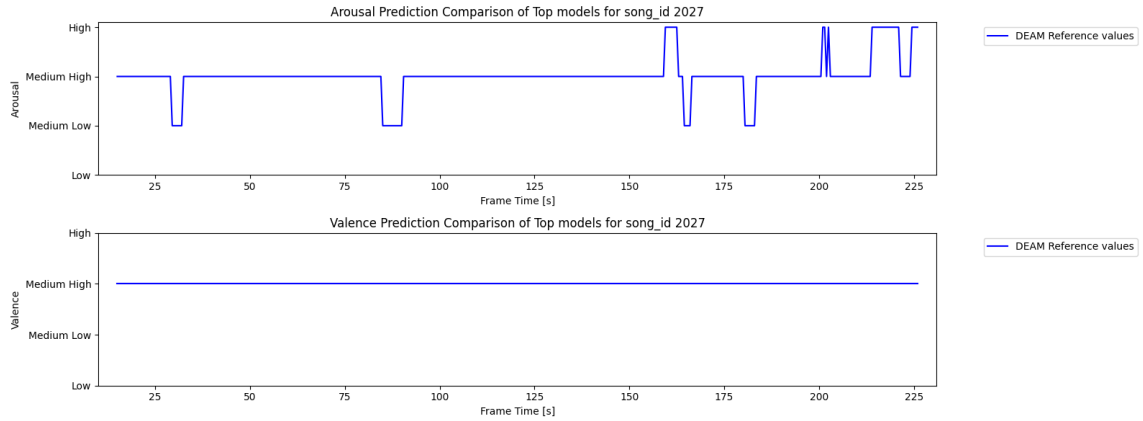


(b) Reference values discretized to High, Medium, and Low categories.

Figure 5.7: Valence and Arousal Reference values discretized to High and Low categories. Visualization for Song # 2012.



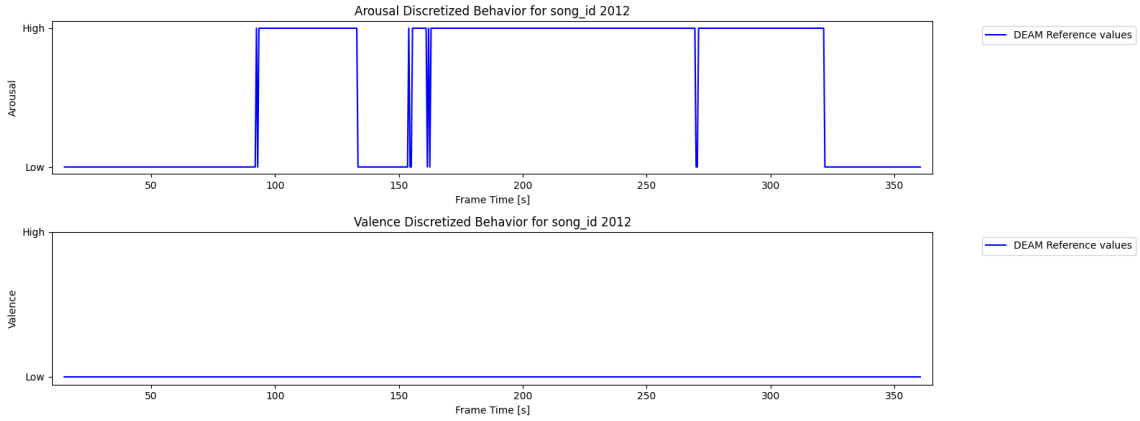
(a) Reference values discretized to High and Low categories.



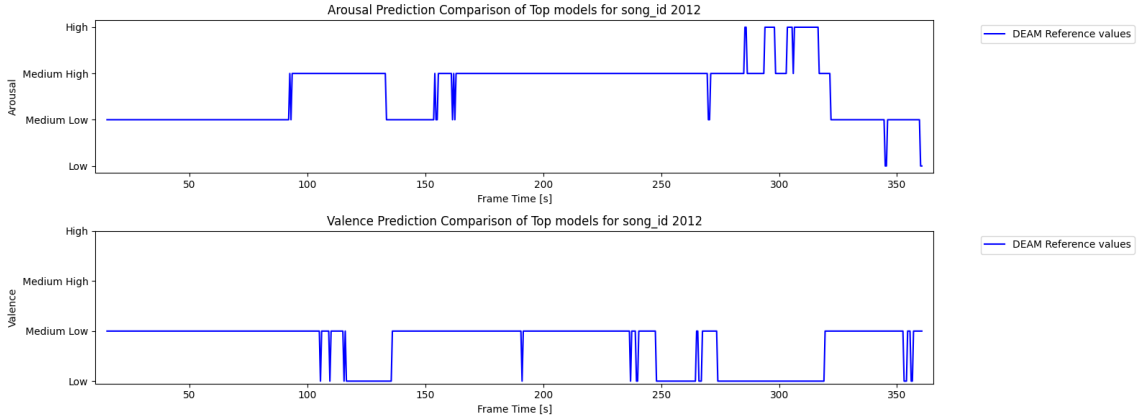
(b) Reference values discretized to High, Medium, and Low categories.

Figure 5.8: Valence and Arousal Reference values discretized to High and Low categories. Visualization for Song # 2027.

Estimation

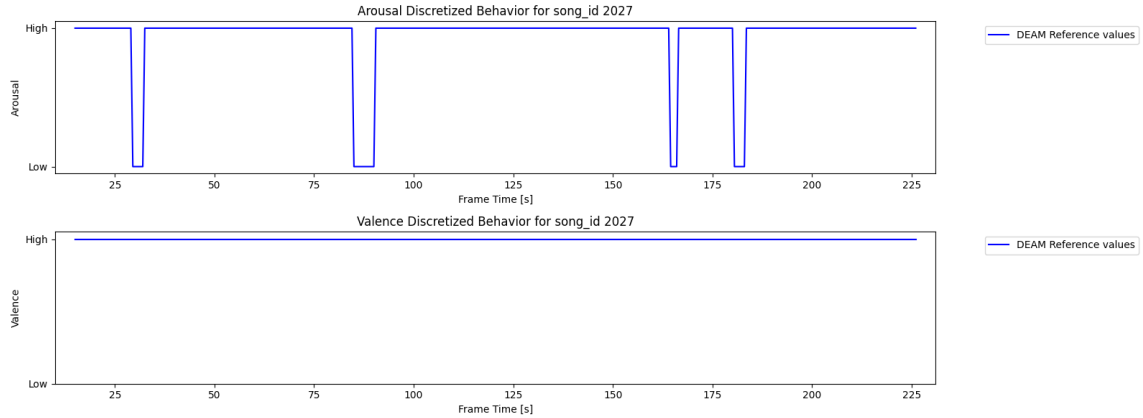


(a) Estimated values discretized to High and Low categories.

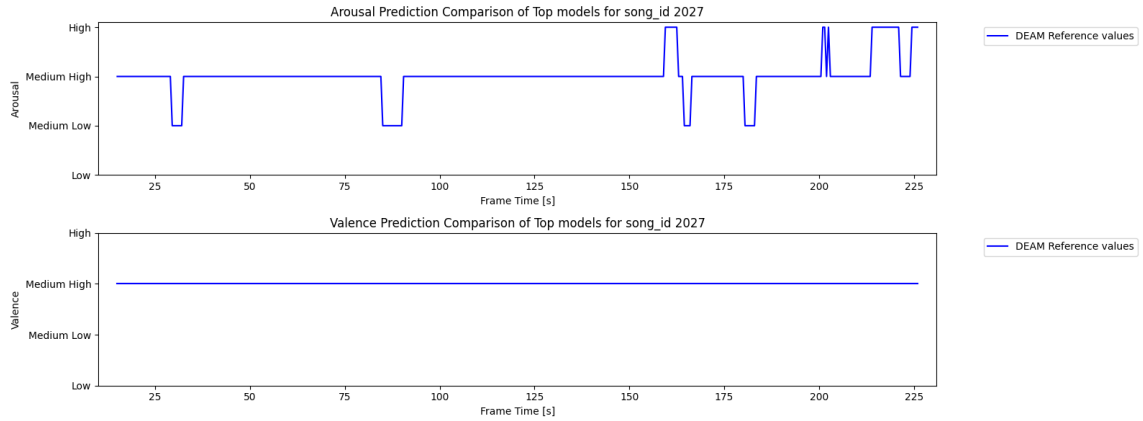


(b) Estimated values discretized to High, Medium, and Low categories.

Figure 5.9: Valence and Arousal Estimated values discretized to High and Low categories. Visualization for Song # 2012.



(a) Estimated values discretized to High and Low categories.



(b) Estimated values discretized to High, Medium, and Low categories.

Figure 5.10: Valence and Arousal Estimated values discretized to High and Low categories. Visualization for Song # 2027.

Among the aforementioned indicators, Tables 5.1 (Song # 2012) and 5.2 (Song # 2027) display the percentage of time spent in each category for the songs used in the demonstration. Notably, even if the prediction values lack accuracy at a dimensional level, they prove useful for a categorical approach, offering a representation of the song's macro emotional behavior. Figure 5.11 illustrates the estimated distribution of emotional tags within each dimension, providing insight into how emotions are grouped across the added categories. For each timestep, the Euclidean distance between the estimated V/A point and the emotions points within the respective category is computed. The objective of this analysis is to generate an indicator based on the amount of time that the estimation closely aligns with a particular emotional tag. This involves assigning a weight value based on the ratio between the amount of instances for a particular emotion against the total instances for the song (song duration) over the average of the distance values obtain from those instances. This means:

$$EmotionRanking = \frac{\#points_{emotion} \times Avg\_distance(estimation, emotion\_position)}{Songlength(\#timesteps)} \quad (5.1)$$

Tables 5.3 (Song # 2012) and 5.4 (Song # 2027) display the top 5 emotions obtained by applying the indicator presented in equation 5.1. The bold values are to remark the similarities on the top 5 emotions between the reference and estimation values for the selected songs.

Table 5.1: Time distribution comparison between the estimated emotional trajectory values and the reference values for Song # 2012.

Song_id 2012	Reference		Estimation	
	Timesteps	%	Timesteps	%
HAHV	-	-	264	38.15
HALV	411	59.39	45	6.50
LAHV	-	-	40	5.78
LALV	281	40.61	343	49.57

Table 5.2: Time distribution comparison between the estimated emotional trajectory values and the reference values for Song # 2027.

Song_id 2027	Reference		Estimation	
	Timesteps	%	Timesteps	%
HAHV	396	93.62	391	92.43
HALV	-	-	12	2.84
LAHV	27	6.38	3	0.71
LALV	-	-	17	4.02

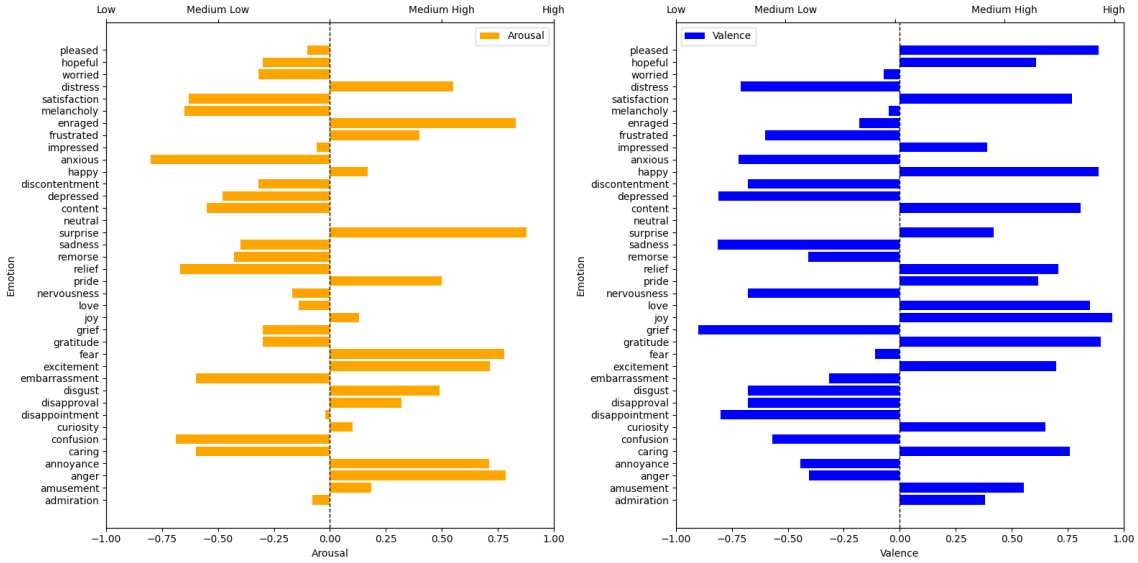


Figure 5.11: Estimated distribution of the emotional tags presented in Subsubsection 2.2.2.1 over each dimension.

Table 5.3: Top 5 emotion percentage distribution for Song # 2012.

Reference		Estimation	
Emotion	%	Emotion	%
Remorse	8,17	**Worried**	15,27
Nervousness	7,42	Neutral	10,74
Worried	7,2	**Remorse**	6,43
Discontentment	7,73	Embarrasement	5,41
Dissapointment	5,84	Melancholy	5,31

Table 5.4: Top 5 emotion percentage distribution for Song # 2027.

Reference		Estimation	
Emotion	%	Emotion	%
Neutral	17,87	**Surprise**	14,89
Amusement	15,8	**Pride**	14,75
Curiosity	12,58	**Amusement**	12,6
Pride	12,32	Excitement	1181
Surprise	9,41	**Curiosity**	11,81

6 | Conclusions

Defining an Emotional System for Music Emotion Recognition is a challenging and demanding research field, as a comprehensive understanding of emotional concepts within the design steps and the structural framework is required. Nevertheless, as presented in the literature review, this complexity allows for the definition and presentation of multiple systems and structures, from simple to complex architectures or user-oriented solutions. The emotional label values are usually linked to specific measurements and experiments and are, therefore, empirically derived. However, there is no consensus on the list of emotions or the corresponding V/A point values. This ambiguity introduces variability into the research field, wherein the positioning of a particular emotion may differ across studies, thereby amplifying the drawbacks of a dynamic model, depending on the granularity of the approach. Despite the absence of standardization, research efforts to generate more of these values and systems based on these values reflect the escalating interest in the field. This chapter provides a general overview of the work carried out in this thesis and a summary of its contributions and proposals. The problems and difficulties are also presented, using them to present future improvements and further research.

6.1 Summary and Conclusions

Through the research and definition process of the proposed Emotional System, it was possible to trace the root area of investigation that fostered the research interests back to the Affective Computing field and to comprehend the context behind the emotional taxonomies available in the literature, along with the benchmarking works and metrics that

facilitate the research comparatives required to develop and implement the final system. Different techniques and approaches within the field were reviewed after selecting the Dimensional Emotion Taxonomy, particularly Russell's Affective model. This process led to the formulation of the research question and the delineation of the field of interest. On the one hand, the concept of the *emotional trajectory* emerged as a means to visualize emotional variation over time and characterize specific pieces of music. On the other hand, implementing a Transformer-based model was considered to estimate these values directly from music pieces and a set of low-level features that represent the behavior of individual songs over time.

Among the main difficulties encountered was the availability of datasets with quality data, which was essential for developing MEVD approaches. Many works had to develop and annotate their datasets due to the need to count with information or details not available in other datasets or not yet considered a research interest to have available data. Another issue identified was the limited amount of data within public datasets. While these datasets facilitate the development of techniques and models within the research field, their restricted volume of values constrains the training of models with broader capabilities. Additionally, the data available in these datasets is known for being annotated by a limited number of subjects, which further restricts the statistical representation of the predictions.

The absence of consensus regarding the structure of MER systems may stem from various factors, including the influence of the chosen emotional taxonomy, the selected datasets, and the objectives of the Emotional System under development. This complexity results in the emergence of multiple system architectures and structures, ranging from simple to intricate designs or user-oriented solutions tailored to meet the specific requirements of the Emotional System. In defining the Emotion Taxonomy, the predominant models in the literature (categorical and dimensional) dictate the primary flow of the Emotional System. This is because it determines whether the problem is approached as a regression or classification task, thus influencing the choice of available datasets and benchmarks to validate the proposed workflow. Additionally, there is a lack of consensus or standardized values concerning emotions in the VA plane. These values are intrinsically subjective and shaped by cultural and linguistic contexts. Consequently, representations often derive from

diverse sources, including crowdsourced annotations and experimental setups for emotional annotation values acquisition.

Emotion Recognition has witnessed significant advancement alongside the development of newer and more sophisticated models to optimize processes and enhance accuracy in these tasks. The introduction of CNN and LSTM approaches represented a significant leap forward in the evolution of MER techniques, addressing limitations inherent in classical methods and expanding the capacity and predictive capabilities of MER systems. Today, models such as Transformer marks a new frontier in system development within the field. However, to fully harness the capabilities of these advanced approaches, additional modalities are being incorporated into the systems, thereby introducing greater complexity to the Domain Definition step. The universality of Music Emotion Prediction lies within the music itself. An ideal Dynamic MER model, characterized by high accuracy and CCC scores, can reconstruct the Emotional Trajectory directly from music audio features. The lower accuracy of the values for the valence dimension exhibited by the implemented models can be attributed to the inherent difficulty in estimating valence based solely on low-level features. Arousal is typically associated with tempo (fast or slow), pitch (high or low), loudness level (high or low), and timbre (bright or soft), whereas the valence dimension is primarily linked to mode (major or minor) and harmony (consonant or dissonant). Yang *et al.* addresses this issue, highlighting the challenges posed by the limited number of features used to asses valence compared to arousal. Additionally, the annotation process presents challenges, as individual differences can significantly impact valence evaluation, potentially leading to varying perceptions of the same song with opposite valence values.

Regarding the evaluation process of the models, it was found that the main selected metrics used in the benchmark dataset were unsuitable for the context, as they did not account for errors in the model evaluation processes, particularly in cases where the prediction resulted in a constant value. Specifically, many models with top RMSE metrics in the valence dimension exhibited poor real accuracy, while the associated CCC metric was very low in comparison. One proposed solution is to modify the metric criteria to give more importance to the CCC metric, as done in the AVEC challenge. Another approach is to implement a CCC-based loss function in the training process, as suggested in works by Sun

et al. [95, 105], as optimizing this loss value directly improves the model's performance. The analysis conducted on the models for evaluating the proposed MRS involved several steps. First, the obtained emotional trajectory was discretized to characterize the signal, followed by applying pattern recognition algorithms to enhance the system recommendation capabilities. Subsequently, the time distribution of the discretized values of each dimension was assessed to reveal the emotional distribution inherent in the musical piece. However, it is important to note that these analyses are directly subject to the base model's accuracy. Introducing a categorized approach atop the discretization characterization could offer alternative means of interacting with the user and showcase the interface capabilities of the recommender system. This approach could allow users to select musical genres or input song fragments to generate a starting point for recommendations, enhancing the system's versatility and user engagement.

6.2 Perspectives for Future Research

There is still plenty of room for improvement and future research in a few key areas. Firstly, addressing the data imbalance within the DEAM dataset is crucial for making the models built on it more robust and widely applicable. Moreover, broadening the scope of training and testing sets by incorporating other benchmark datasets could directly impact model accuracy and behavior. Another promising avenue is the implementation of a CCC loss function into the model's training process. This could improve the evaluation metrics used to assess the predictive performance of the models. Additionally, further exploration of diverse model implementations and extended training processes is warranted to explore their potential efficacy. Regarding the original proposal for the Recommender System, which envisioned a potential integration with the Spotify API for streamlined song excerpt retrieval, enabling personalized music recommendations based on user input. However, the prohibitive usage restrictions imposed by Spotify's Developer Policy limit access to only genre tags and high-level features, diverging from the initial focus of the project. Also, the usage of this data is not allowed for machine learning projects, which further complicates the envisioned integration process.

Bibliography

[1] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, Association for Computing Machinery, Oct 2010.

[2] M. Larson, S. Papadopoulos, M. Soleymani, Y.-H. Yang, G. Jones, R. Sutcliffe, C. Hauff, M. Eskevich, M. Riegler, J. Poignant, and et al., "Mediaeval 2015 - Multimedia Benchmark Workshop." https://ceur-ws.org/Vol-1436/, Sep 2015.

[3] A. Tzacheva, D. Schlingmann, and K. Bell, "Automatic detection of emotions with music files," *International Journal of Social Network Mining (IJSNM)*, Jan. 2012.

[4] G. Paltoglou and M. Thelwall, "Seeing Stars of Valence and Arousal in Blog Posts," *IEEE Transactions on Affective Computing*, Jan. 2013.

[5] G. N. Coutinho, A. d. V. Lima, J. Yugoshi, M. I. d. M. Junior, M. P. S. Gôlo, and R. M. Marcacini, "Multimodal Audio Emotion Recognition with Graph-based Consensus Pseudolabeling," in *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, SBC, Sept. 2023.

[6] Y.-S. Seo and J.-H. Huh, "Automatic Emotion-Based Music Classification for Supporting Intelligent IoT Applications," *Electronics*, Feb. 2019.

[7] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLOS ONE*, Mar. 2017.

[8] E. Y. Koh, K. W. Cheuk, K. Y. Heung, K. R. Agres, and D. Herremans, "MERP: A Music Dataset with Emotion Ratings and Raters' Profile Information," *Sensors*, Jan. 2023.

[9] D. Bogdanov, X. Lizarraga-Seijas, P. Alonso-Jiménez, and X. Serra, "MusAV: A dataset of relative arousal-valence annotations for validation of audio models," in *International Society for Music Information Retrieval Conference (ISMIR 2022)*, Dec. 2022.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017.

[11] J. Dharmapriya, L. Dayarathne, T. Diasena, S. Arunathilake, N. Kodikara, and P. Wijesekera, "Music Emotion Visualization through Colour," in *2021 International Conference on Electronics, Information, and Communication (ICEIC)*, Jan. 2021.

[12] E. Coutinho, A. Alshukri, J. de Berardinis, and C. Dowrick, "POLYHYMNIA Mood - Empowering people to cope with depression through music listening," in *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '21 Adjunct, Association for Computing Machinery, Sept. 2021.

[13] Y. O. Medina, J. R. Beltrán, C. Sanz, and S. Baldassarri, "Dimensional Emotion Prediction through Low-Level Musical Features," in *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound*, AM'19, Association for Computing Machinery, May 2019.

[14] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal Transformer Fusion for Continuous Emotion Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2020.

[15] A. Rodriguez Aguiñaga, L. Muñoz Delgado, V. R. López-López, and A. Calvillo Téllez, "EEG-Based Emotion Recognition Using Deep Learning and M3GP," *Applied Sciences*, Jan. 2022.

[16] V. Chaturvedi, A. B. Kaur, V. Varshney, A. Garg, G. S. Chhabra, and M. Kumar, "Music mood and human emotion recognition based on physiological signals: a systematic review," *Multimedia Systems*, Feb. 2022.

[17] R. W. Picard, *Affective Computing*. The MIT Press, July 2000.

[18] L. Yang, Z. Shen, J. Zeng, X. Luo, and H. Lin, "COSMIC: Music emotion recognition combining structure analysis and modal interaction," *Multimedia Tools and Applications*, July 2023.

[19] L. Feng, C. Cheng, M. Zhao, H. Deng, and Y. Zhang, "EEG-Based Emotion Recognition Using Spatial-temporal Graph Convolutional LSTM with Attention Mechanism," *IEEE Journal of Biomedical and Health Informatics*, Aug. 2022.

[20] R. Chaudhary and R. A. Jaswal, "A Review of Emotion Recognition Based on EEG using DEAP Dataset," *International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET)*, June 2021.

[21] F. Galvão, S. M. Alarcão, and M. J. Fonseca, "Predicting Exact Valence and Arousal Values from EEG," *Sensors*, May 2021.

[22] H. Tran, T. Le, A. Do, T. Vu, S. Bogaerts, and B. Howard, "Emotion-Aware Music Recommendation," *Proceedings of the AAAI Conference on Artificial Intelligence*, Sept. 2023.

[23] V. Moscato, A. Picariello, and G. Sperlí, "An Emotional Recommender System for Music," *IEEE Intelligent Systems*, Sept. 2021.

[24] F. Andayani, "Investigating the Impacts of LSTM-Transformer on Classification Performance of Speech Emotion Recognition," 2022.

[25] P. Singh Tomar, K. Mathur, and U. Suman, "Unimodal approaches for emotion recognition: A systematic review," *Cognitive Systems Research*, Jan. 2023.

[26] K. Ezzameli and H. Mahersia, "Emotion recognition from unimodal to multimodal analysis: A review," *Information Fusion*, Nov. 2023.

[27] R. V. Aranha, C. G. Corrêa, and F. L. S. Nunes, "Adapting Software with Affective Computing: A Systematic Review," *IEEE Transactions on Affective Computing*, Oct. 2021.

[28] P. Knees, M. Schedl, and M. Goto, "Intelligent User Interfaces for Music Discovery," *Transactions of the International Society for Music Information Retrieval*, Oct. 2020.

[29] J. S. Downie, "The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Future," *Computer Music Journal*, June 2004.

[30] X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gómez Gutiérrez, F. Gouyon, P. Herrera Boyer, S. Jordà Puig, O. Paytuvi, G. Peeters, J. Schlüter, H. Vinet, and G. Widmer, *Roadmap for Music Information ReSearch*. The MIReS Consortium, 2013.

[31] J. S. Gomez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gomez, "Music Emotion Recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, Nov. 2021.

[32] IFPI, "Global music report 2023." https://ifpi-website-cms.s3.eu-west-2.amazonaws.com/GMR_2023_State_of_the_Industry_ee2ea600e2.pdf. Accessed: 2023-07-25.

[33] Spotify, "About spotify." https://newsroom.spotify.com/company-info/. Accessed: 2023-07-25.

[34] M. Schedl, E. Gomez, and J. Urbano, "Music Information Retrieval: Recent Developments and Applications," *Foundations and Trends in Information Retrieval*, Sept. 2014.

[35] D. Han, Y. Kong, J. Han, and G. Wang, "A survey of music emotion recognition," *Frontiers of Computer Science*, Jan. 2022.

[36] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Systems*, July 2018.

[37] J. Yang, "A Novel Music Emotion Recognition Model Using Neural Network Technology," *Frontiers in Psychology*, 2021.

[38] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, Association for Computing Machinery, Mar. 2010.

[39] D. Perera, M. Rajaratne, S. Arunathilake, K. Karunanayaka, and B. Liyanage, "A Critical Analysis of Music Recommendation Systems and New Perspectives," *International Conference on Human Interaction and Emerging Technologies (IHIET)*, Apr. 2020.

[40] B. Vad, D. Boland, J. Williamson, R. Murray-Smith, and P. B. Steffensen, "Design and evaluation of a probabilistic music projection interface," *International Society for Music Information Retrieval Conference*, 2015.

[41] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi, "Current challenges and visions in music recommender systems research," *International Journal of Multimedia Information Retrieval*, vol. 7, June 2018.

[42] K. R. Scherer, "What are emotions? And how can they be measured?," *Social Science Information*, Dec. 2005.

[43] J. J. Gross and L. F. Barrett, "Emotion Generation and Emotion Regulation: One or Two Depends on Your Point of View," *Emotion review*, Jan. 2011.

[44] R. E. S. Panda, *Emotion-based Analysis and Classification of Audio Music*. PhD thesis, Universidade d Coimbra, May 2019.

[45] Y.-H. Yang and H. H. Chen, *Music emotion recognition*. CRC Press, Inc., 2011.

[46] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, Jan. 2011.

[47] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, and S. K. Saha, "Recognition of emotion in music based on deep convolutional neural network," *Multimedia Tools and Applications*, Jan. 2020.

[48] R. Orješek, R. Jarina, and M. Chmulik, "End-to-end music emotion variation detection using iteratively reconstructed deep features," *Multimedia Tools and Applications*, Feb. 2022.

[49] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, Dec. 1980.

[50] M. Kowalska and M. Wróbel, "Basic Emotions," in *Encyclopedia of Personality and Individual Differences*, Springer International Publishing, July 2017.

[51] T. Eerola and J. K. Vuoskoski, "A Review of Music and Emotion Studies: Approaches, Emotion Models, and Stimuli," *Music Perception: An Interdisciplinary Journal*, 2013.

[52] J. J. Deng, C. H. C. Leung, A. Milani, and L. Chen, "Emotional States Associated with Music: Classification, Prediction of Changes, and Consideration in Recommendation," *ACM Transactions on Interactive Intelligent Systems*, Mar. 2015.

[53] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, Mar. 2011.

[54] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, 1992.

[55] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, Feb. 1999.

[56] P. Ekman and D. Cordaro, "What is Meant by Calling Emotions Basic," *Emotion Review*, Sept. 2011.

[57] K. Hevner, "Expression in music: a discussion of experimental studies and theories," *Psychological Review*, 1935.

[58] K. Hevner, "Experimental Studies of the Elements of Expression in Music," *The American Journal of Psychology*, Apr. 1936.

[59] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and M. Yang, "Multimodal Continuous Emotion Recognition with Data Augmentation Using Recurrent Neural Networks," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, AVEC'18, Association for Computing Machinery, Oct. 2018.

[60] N. N. Vempala and F. A. Russo, "Modeling Music Emotion Judgments Using Machine Learning Methods," *Frontiers in Psychology*, Jan. 2018.

[61] R. E. Thayer, *The Biopsychology of Mood and Arousal*. Oxford University Press USA, 1989.

[62] Y. Ma, X. Li, M. Xu, J. Jia, and L. Cai, "Multi-scale Context Based Attention for Dynamic Music Emotion Prediction," MM '17, Association for Computing Machinery, Oct. 2017.

[63] J. Grekow, "Music Emotion Maps in Arousal-Valence Space," in *Computer Information Systems and Industrial Management*, Springer International Publishing, 2016.

[64] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller, "Discriminatively Trained Recurrent Neural Networks for Continuous Dimensional Emotion Recognition from Audio," July 2016.

[65] A. Warmbrodt, R. Timmers, and R. Kirk, "The emotion trajectory of self-selected jazz music with lyrics: A psychophysiological perspective," *Psychology of Music*, May 2022.

[66] M. Soleymani, A. Aljanaki, and Y.-H. Yang, "DEAM: MediaEval Database for Emotional Analysis in Music," Apr. 2018.

[67] M. Soleymani and M. N. Caro, "The MediaEval 2013 Brave New Task: Emotion in Music," Oct. 2013. MediaEval 2013 Workshop.

[68] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Emotion in Music Task at MediaEval 2014," Oct. 2014. MediaEval 2014 Workshop.

[69] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Emotion in Music Task at MediaEval 2015," Sept. 2015. MediaEval 2015 Workshop.

[70] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Emotion in Music task: lessons learned," Oct. 2016. MediaEval 2016 Workshop.

[71] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, Mar. 1994.

[72] M. Soleymani, A. Aljanaki, Y.-H. Yang, M. N. Caro, F. Eyben, K. Markov, B. W. Schuller, R. Veltkamp, F. Weninger, and F. Wiering, "Emotional analysis of music: A comparison of methods," in *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, Association for Computing Machinery, 2014.

[73] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, ACM, Oct. 2013.

[74] N. Kumar, R. Gupta, T. Guha, C. Vaz, M. V. Segbroeck, J. Kim, and S. S. Narayanan, "Affective Feature Design and Predicting Continuous Affective Dimensions from Music," Oct. 2014. MediaEval 2014 Workshop.

[75] L. I.-K. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, 1989.

[76] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2015: The 5th International Audio/Visual Emotion Challenge and Workshop," in *Proceedings of the 23rd ACM international conference on Multimedia*, MM '15, Association for Computing Machinery, Oct. 2015.

[77] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge," AVEC '16, Association for Computing Machinery, Nov. 2016.

[78] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, Association for Computing Machinery, Oct. 2017.

[79] K. E. Yan and D. Herremans, "Music emotion recognition with profile information." https://www.kaggle.com/datasets/kohenyan/music-emotion-recognition-with-profile-information.

[80] D. Herremans, "Music emotion recognition with profile information (merp)." https://github.com/dorienh/MERP, 2022.

[81] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The Distress Analysis Interview Corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), May 2014.

[82] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, and M. Pantic, "SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Mar. 2021.

[83] R. Panda, R. Malheiro, and R. P. Paiva, "Audio Features for Music Emotion Recognition: A Survey," *IEEE Transactions on Affective Computing*, Jan. 2023.

[84] F. Eyben, F. Weninger, M. Wollmer, and B. Schuller, "open-Source Media Interpretation by Large feature-space Extraction," 2014.

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

[86] Y. O. Medina, J. R. Beltrán, and S. Baldassarri, "Emotional classification of music using neural networks with the MediaEval dataset," *Pers Ubiquit Comput*, Aug. 2022.

[87] T. Krols, Y. Nikolova, and N. Oldenburg, "Multi-Modality in Music: Predicting Emotion in Music from High-Level Audio Features and Lyrics," Feb. 2023.

[88] S. O. Clement Allognon, A. de S. Britto, and A. L. Koerich, "Continuous Emotion Recognition via Deep Convolutional Autoencoder and Support Vector Regressor," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jan. 2020.

[89] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, "Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition," in *Proceedings of the 14th Sound and Music Computing Conference 2017*, Aalto University, June 2017.

[90] Y. Dong, X. Yang, X. Zhao, and J. Li, "Bidirectional Convolutional Recurrent Sparse Network (BCRSN): An Efficient Model for Music Emotion Recognition," *IEEE Transactions on Multimedia*, Dec. 2019.

[91] X. Liu, Q. Chen, X. Wu, Y. Liu, and Y. Liu, "CNN based music emotion classification," Apr. 2017. arXiv:1704.05665 [cs].

[92] R. Orjesek, R. Jarina, M. Chmulik, and M. Kuba, "DNN Based Music Emotion Recognition from Raw Audio Signal," in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, Apr. 2019.

[93] X. Li, J. Tian, M. Xu, Y. Ning, and L. Cai, "DBLSTM-based multi-scale fusion for dynamic emotion prediction in music," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, July 2016.

[94] H. Liu, Y. Fang, and Q. Huang, "Music Emotion Recognition Using a Variant of Recurrent Neural Network," in *Proceedings of the 2018 International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA 2018)*, Atlantis Press, Jan. 2019.

[95] L. Sun, Z. Lian, J. Tao, B. Liu, and M. Niu, "Multi-modal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism," in *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, MuSe'20, Association for Computing Machinery, Oct. 2020.

[96] L. Zhang, X. Yang, Y. Zhang, and J. Luo, "Dual Attention-Based Multi-Scale Feature Fusion Approach for Dynamic Music Emotion Recognition," 2023.

[97] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files," *IEEE Access*, 2022.

[98] J. Vazquez-Rodriguez, G. Lefebvre, J. Cumin, and J. L. Crowley, "Transformer-Based Self-Supervised Learning for Emotion Recognition," in *2022 26th International Conference on Pattern Recognition (ICPR)*, Aug. 2022.

[99] Y. Agrawal, R. G. R. Shanker, and V. Alluri, "Transformer-Based Approach Towards Music Emotion Recognition from Lyrics," in *Advances in Information Retrieval*, Lecture Notes in Computer Science, Springer International Publishing, 2021.

[100] Y. He, L. Sun, Z. Lian, B. Liu, J. Tao, M. Wang, and Y. Cheng, "Multimodal Temporal Attention in Sentiment Analysis," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, MuSe' 22, Association for Computing Machinery, Oct. 2022.

[101] L. Christ, S. Amiriparian, A. Baird, P. Tzirakis, A. Kathan, N. Müller, L. Stappen, E.-M. Meßner, A. König, A. Cowen, E. Cambria, and B. W. Schuller, "The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, Association for Computing Machinery, Oct. 2022.

[102] Z. Zhao, Y. Wang, G. shen, Y. Xu, and J. Zhang, "TDFNet: Transformer-based Deep-scale Fusion Network for Multimodal Emotion Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[103] L. Guo, L. Wang, J. Dang, Y. Fu, J. Liu, and S. Ding, "Emotion Recognition With Multimodal Transformer Fusion Framework Based on Acoustic and Lexical Information," *IEEE MultiMedia*, Apr. 2022.

[104] J.-H. Hsu and C.-H. Wu, "Applying Segment-Level Attention on Bi-Modal Transformer Encoder for Audio-Visual Emotion Recognition," *IEEE Transactions on Affective Computing*, Oct. 2023.

[105] H. Sun, Y.-W. Chen, and L. Lin, "TensorFormer: A Tensor-Based Multimodal Transformer for Multimodal Sentiment Analysis and Depression Detection," *IEEE Transactions on Affective Computing*, Oct. 2023.

[106] J. Grekow, "Audio features dedicated to the detection of arousal and valence in music recordings," in *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, July 2017.

[107] A. Aljanaki, F. Wiering, and R. C. Veltkamp, "MediaEval 2015: A Segmentation-based Approach to Continuous Emotion Tracking."

[108] R. Gupta and S. Narayanan, "Predicting Affect in Music Using Regression Methods on Low Level Features." MediaEval 2015 Workshop.

[109] T. Pellegrini and V. Barrière, "Time-continuous Estimation of Emotion in Music with Recurrent Neural Networks." MediaEval 2015 Workshop.

[110] E. Coutinho, G. Trigeorgis, S. Zafeiriou, and B. Schuller, "Automatically Estimating Emotion in Music with Deep Long-Short Term Memory Recurrent Neural Networks." MediaEval 2015 Workshop.

[111] M. Xu, X. Li, H. Xianyu, J. Tian, F. Meng, and W. Chen, "Multi-scale Approaches to the MediaEval 2015 "Emotion in Music" Task." MediaEval 2015 Workshop.

[112] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "ESSENTIA: an Audio Analysis Library for Music Information Retrieval," *Proceedings - 14th International Society for Music Information Retrieval Conference*, Nov. 2013.

[113] "DEAM dataset - Database for Emotional Analysis of Music." https://cvml.unige.ch/databases/DEAM/.

A | Transformer Model Code Snippet

```
class MultiHeadAttention(layers.Layer):
def __init__(self, d_model, num_heads):
    super(MultiHeadAttention, self).__init__()

    self.num_heads = num_heads # number of attention heads to use

    assert d_model % self.num_heads == 0
    self.depth = d_model // self.num_heads

    self.d_model = d_model # dimensionality of the model
    self.query_dense = layers.Dense(d_model)
    self.key_dense = layers.Dense(d_model)
    self.value_dense = layers.Dense(d_model)
    self.dense = layers.Dense(d_model)

def split_heads(self, x, batch_size):
    x = tf.reshape(x, (batch_size, -1, self.num_heads, self.depth))
    return tf.transpose(x, perm=[0, 2, 1, 3])

def scaled_dot_product_attention(self, q, k, v, mask=None):
    matmul_qk = tf.matmul(q, k, transpose_b=True)
    dk = tf.cast(tf.shape(k)[-1], tf.float32)
    scaled_attention_logits = matmul_qk / tf.math.sqrt(dk)

    if mask is not None:
        scaled_attention_logits += (mask * -1e9)

    attention_weights = tf.nn.softmax(scaled_attention_logits, axis=-1)
    output = tf.matmul(attention_weights, v)
```

```
        return output, attention_weights

    def call(self, q, k, v, mask):
        batch_size = tf.shape(q)[0]

        q = self.query_dense(q)
        k = self.key_dense(k)
        v = self.value_dense(v)

        q = self.split_heads(q, batch_size)
        k = self.split_heads(k, batch_size)
        v = self.split_heads(v, batch_size)

        scaled_attention, attention_weights = self.scaled_dot_product_attention(q, k, v, mask)

        scaled_attention = tf.transpose(scaled_attention, perm=[0, 2, 1, 3])
        concat_attention = tf.reshape(scaled_attention, (batch_size, -1, self.d_model))

        output = self.dense(concat_attention)
        return output, attention_weights
```