UNIVERSIDADE D
COIMBRA

Tomás Gomes Ferreira

Music Emotion Variation Detection: A Comparative Analysis of Variable and Fixed-sized Window Techniques in Machine Learning and Deep Learning

Dissertation in the context of Master in Data Science and Engineering, advised by Professors Renato Panda, Rui Pedro Paiva and Pedro Louro and presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

July 2024

DEPARTAMENTO DE
ENGENHARIA INFORMÁTICA
FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

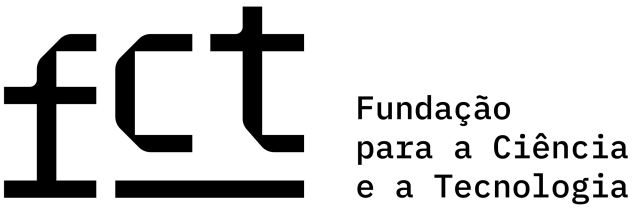
Tomás Gomes Ferreira

Music Emotion Variation Detection: A Comparative Analysis of Variable and Fixed-sized Window Techniques in Machine Learning and Deep Learning

Dissertation in the context of Master in Data Science and Engineering, advised by Professors Renato Panda, Rui Pedro Paiva and Pedro Louro and presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

July 2024

This work is funded by FCT - Foundation for Science and Technology, I.P., within the scope of the projects: MERGE - DOI: 10.54499/PTDC/CCI-COM/3171/2021 financed with national funds (PIDDAC) via the Portuguese State Budget; and project CISUC - UID/CEC/00326/2020 with funds from the European Social Fund, through the Regional Operational Program Centro 2020. Renato Panda was supported by Ci2 - FCT UIDP/05567/2020.



Acknowledgements

Reflecting on a year filled with deep research, numerous experiments, and endless hours spent at my computer striving to produce my finest work, I'm reminded of all the individuals who have been by my side during this demanding journey.

I want to extend my sincere thanks to my advisors, Prof. Renato Panda, Prof. Rui Pedro Paiva and Prof.Pedro Louro for their unwavering support and valuable insights. Their readiness to assist, insightful guidance, and constructive feedback significantly eased my journey through the vast scope of ideas and literature relevant to this project. Their detailed and thoughtful critique has immensely contributed to making my experience both enriching and enjoyable.

I am immensely grateful to my family, especially my mother, Maria and father, Amílcar, for their unwavering care, patience, and unconditional support. They have been my rock, always there to lift me up and never letting me give up on completing this project. Their support has been invaluable to me.

I also want to highlight the exceptional support I received from Hugo. Throughout the development of this project, they directly addressed all my concerns and frustrations. They went out of their way to ensure I had moments of enjoyment and relaxation, and they consistently reminded me of my objectives, even when achieving them felt out of reach.

I want to thank all my friends and family for their supportive words during our conversations.

I am deeply grateful to the University of Coimbra and sincerely thank all the faculty and colleagues who played a significant role in my academic journey. Moreover, I'd like to acknowledge the support from the MERGE project (Music Emotion Recognition - Next Generation, PTDC/CCI-COM/3171/2021, DOI: 10.54499/PTDC/CCI-COM/3171/2021), funded by the Fundação para a Ciência e Tecnologia (FCT), which provided a valuable context for my work.

I could not have completed this work without the support of everyone involved in this journey. My heartfelt thanks go out to each of you for your support and the precious memories I will hold dear for years to come. Thank you for believing in me when I doubted myself, and above all, thank you for your constant presence and for helping me grow as an individual.

Abstract

The rise of digital music streaming platforms has led to a growing interest in Music Emotion Recognition (MER). Within MER, Music Emotion Variation Detection (MEVD) is a topic of increasing relevance, focusing on the analysis of emotion variation throughout songs, rather than focusing only on the dominant emotion. Traditional approaches in this field typically involve feature engineering to classify song emotions, but recent advancements in deep learning methodologies using neural networks have shown great promise in achieving the same objective. However, challenges persist, such as the reliance on small or low-quality datasets and unsuitable features for emotion classification, which can limit the accuracy and effectiveness of these models.

This study extends the work of Panda and Paiva from 2011 on MER by exploring the potential of the All-In-One tool for music segmentation. The primary objective is to assess whether this tool, which aims to eliminate issues associated with small windows or segmentation methods prone to artifacts and boundary effects, can enhance emotion recognition performance. Extensive experiments were conducted using both Deep Learning (DL) and Machine Learning (ML) approaches in combination with the All-in-One structural segmentation tool. The research evaluates the impact of dynamic window sizes created by this tool compared to traditional fixed-size windows of 1.5 seconds.

The experiments yielded promising results for the variable-size windows, with the best F1-Score outcomes being **53.17%** for the SVM approach and **36.94%** for the DL approach. Although these results do not surpass the current state-of-the-art benchmarks, they highlight the potential of variable-sized windows through structural segmentation in improving emotion classification in music.

This work contributes by demonstrating the potential benefits of using the All-in-One tool for dynamic segmentation, which could lead to improvements in the MEVD field. However, as one of the first structural segmentation tools applied in this context, the All-in-One tool shows significant promise but requires further refinement to fully exploit its capabilities. This study's findings contribute to developing more accurate and reliable methods for MER, potentially impacting future research and applications in the field.

Keywords

Music Emotion Recognition, Music Emotion Variation Detection, Music Information Retrieval, Audio Analysis, Music Segmentation.

Resumo

O aumento das plataformas de streaming de música digital levou a um interesse crescente no Reconhecimento de Emoção em Música (REM). No âmbito do REM, Deteção de Variação de Emoção em Música (DVEM) é um tópico de crescente relevância, centrando-se na análise da variação da emoção ao longo das canções, em vez de se concentrar apenas na emoção dominante. As abordagens tradicionais neste domínio envolvem tipicamente a engenharia de caraterísticas para classificar as emoções das canções, mas os recentes avanços nas metodologias de aprendizagem profunda utilizando redes neuronais têm-se revelado muito promissores para alcançar o mesmo objetivo. No entanto, persistem desafios, como a dependência de conjuntos de dados pequenos ou de baixa qualidade e caraterísticas inadequadas para a classificação de emoções, o que pode limitar a precisão e a eficácia desses modelos.

O presente estudo alarga o trabalho do Panda and Paiva de 2011 sobre o REM, explorando o potencial da ferramenta All-in-One para a segmentação musical. O objetivo principal é avaliar se esta ferramenta, que visa eliminar problemas associados a pequenas janelas ou métodos de segmentação propensos a artefactos e efeitos de limite, pode melhorar o desempenho do reconhecimento de emoções. Foram realizadas experiências exaustivas utilizando abordagens de Aprendizagem Profunda (AP) e de Aprendizagem Automática (AA) em combinação com a ferramenta de segmentação estrutural All-in-One. A investigação avalia o impacto dos tamanhos de janela dinâmicos criados por esta ferramenta em comparação com as janelas tradicionais de tamanho fixo de 1,5 segundos.

As experiências produziram resultados promissores para as janelas de tamanho variável, com os melhores resultados de F1-Score a serem **53.17%** para a abordagem SVM e **36.94%** para a abordagem DL. Embora estes resultados não ultrapassem os actuais benchmarks de última geração, destacam o potencial da segmentação dinâmica para melhorar a classificação de emoções na música.

Os contributos deste trabalho incluem a demonstração dos potenciais benefícios da utilização da ferramenta All-in-One para a segmentação dinâmica, o que poderá conduzir a melhorias no domínio do MEVD. No entanto, sendo uma das primeiras ferramentas de segmentação estrutural aplicadas neste contexto, a ferramenta All-in-One mostra-se bastante promissora, mas necessita de ser aperfeiçoada para que as suas capacidades sejam plenamente realizadas. As conclusões do estudo contribuem para o desenvolvimento de métodos mais precisos e fiáveis para a MER, com potencial impacto na investigação e nas aplicações futuras neste domínio.

Palavras-Chave

Reconhecimento de Emoções na Música, Detecção de Variação de Emoção na Música, Recuperação de Informação Musical, Análise de Áudio, Segmentação Musical.

Contents

Acronyms

30x3-fold CV 3-fold cross-validation experiment with 30 repetitions.

4QAED 4-Quadrant Audio Emotion Dataset.

A/V Arousal/Valence.

ABC Artificial Bee Colony.

Adam Adaptive Moment Estimation.

ANN Artificial Neural Network.

AUC Area Under the ROC Curve.

BCRSN Bidirectional Convolutional Recurrent Sparse Network.

BP BackPropagation.

CNN Convolutional Neural Network.

CRNN Convolution Recurrent Neural Network.

dB Decibels.

DBN Dynamic Bayesian Networks.

DEAM MediaEval Database for Emotional Analysis of Music.

DiNA Dilated Neighborhood Attention.

DL Deep Learning.

DNN Dense Neural Network.

F0 Fundamental Frequency.

FCN Fully Convolutional Network.

FFS Forward Feature Selection.

FFT Fast Fourier Transform.

GEMS Geneva Emotional Musical Scale.

GP Gaussian Processes.

GPU Graphics Processing Unit.

GRU Gated Recurrent Unit.

KNN K-nearest Neighbor.

LLMs Large Language Models.

LSTM Long-Short Term Memory.

MAE Mean Absolute Error.

Marsyas Music Analysis Retrieval and Synthesis for Audio Signals.

MEM MediaEval Emotion in Music.

MER Music Emotion Retrieval.

MERGE Music Emotion Recognition - Next Generation.

MEVD Music Emotion Variation Detection.

MFCC Mel-frequency cepstrum coefficients.

MIR Music Information Retrieval.

MIREX Music Information Retrieval eXchange.

ML Machine Learning.

MSD Million Song Dataset.

MSE Mean Squared Error.

NA Neighborhood Attention.

NN Neural Network.

PCA Principal Component Analysis.

RBF Radial Basis Function.

ReLU Rectified Linear Unit.

RMSE Root Mean Squared Error.

RNN Recurrent Neural Network.

RWC Real World Computing.

SCAE Sparse Convolutional Autoencoder.

SGD Stochastic Gradient Descent.

SVC Support Vector Classification.

SVM Support Vector Machine.

SVR Support Vector Regression.

WHBR Weighted Hybrid Binary Representation.

ZCR Zero Crossing Rate.

List of Figures

List of Tables

Chapter 1

Introduction

Music Emotion Retrieval (MER) is a field of Music Information Retrieval (MIR) that focuses on identifying and categorizing the emotional content of a musical piece. The static MER approach views the entire musical composition as a single entity and assigns it one or more emotional labels based on its overall dominant emotional characteristics.

Music Emotion Variation Detection (MEVD) is a topic of MER that identifies emotional changes in a music piece over time. It assigns emotional labels to capture the variations as they unfold, providing a more nuanced understanding of a music piece's emotional landscape.

The study of MER is an intriguing pursuit that delves into the complex relationship between sound and emotion. As an interdisciplinary field merging musicology, computer science, and psychology, MER aims to decode the emotional nuances hidden within musical compositions. By utilizing advanced technologies such as Machine Learning (ML) and signal processing, researchers strive to develop systems that can comprehend, classify, and retrieve the emotional subtleties embedded within melodies, providing a nuanced understanding of the profound connection between music and human emotion.

This exploration into the intersection of art and science holds enormous potential for music enthusiasts and applications in various fields, including personalized music recommendations, affective computing, and the creation of emotionally resonant audiovisual experiences.

Let us embark on a melodic journey through the landscape of MER and uncover the intricate tapestry of emotions woven into the fabric of our favorite tunes, transcending the auditory experience into a realm of profound emotional understanding.

1.1 Problem and Motivation

Exploring the complexities of MER unveils a rich landscape of challenges, yet within each challenge lies an opportunity for progress and innovation.

One of the significant challenges in MER is the subjective nature of emotion classification, which can introduce ambiguity and make it challenging to categorize songs effectively. The variability between songs and sudden emotional shifts within the same track further complicate accurate classification. The study focuses on perceived emotion to mitigate this subjectivity, reducing personal biases in the classification process. Emotion annotations are determined by an absolute majority consensus among multiple annotators, ensuring that the classification reflects a collective agreement and enhancing the reliability of the results.

The MER's lack of emotionally relevant features further complicates accurately identifying emotions within musical pieces. Moreover, the scarcity of comprehensive datasets with high-quality annotations is another hurdle, mainly when aiming for larger datasets essential for testing Deep Learning (DL) approaches.

Creating suitable datasets is an uphill task that involves meticulous annotation efforts and ensuring the inclusion of diverse emotional contexts across various music genres. Limited dataset sizes pose a hindrance, as numerous studies rely on smaller datasets, impeding the implementation of DL methods that thrive on ample data.

The choice of appropriate window size for audio segment classification significantly influences the accuracy of emotion representation. A tiny window might overlook subtle emotional nuances in the music, leaving an inadequate reflection of the overall emotional context. Conversely, a vast window can encompass multiple emotional states within a single segment, amalgamating and misinterpreting distinct emotional characteristics.

Segmentation tools come into play to address this challenge and enhance segmentation. These tools detect boundary changes within the music, identifying shifts or transitions between different sections. By incorporating these tools, the segmentation process becomes more sophisticated, allowing for a nuanced representation of emotions. Applying tools like All-in-One [Kim and Nam, 2023] and DeepChorus [He et al., 2022] is particularly beneficial for researchers, who may encounter challenges in defining an optimal window size. The ability of these tools to identify boundary changes helps with the segmentation align with the natural divisions and transitions in the music, resulting in a more accurate capture of emotional flow and contributing significantly to achieving a more granular and precise classification of emotions within individual audio segments, improving the outcomes of MEVD studies where defining an appropriate window size is complex.

1.2 Objectives and Approaches

The main goal of this project is to advance MER and MEVD by studying classical and DL methods. The objective is to improve the accuracy and consistency of emotional analysis in music by combining these approaches with segmentation tools.

Two classic approaches have been identified when studying emotional information in music. The first approach involves using static windows to analyze fixed segments of audio, for example, 1 second long, to extract emotional information. The second approach involves dividing the composition into distinct parts and analyzing each segment's specific characteristics using tools such as All-in-One [Kim and Nam, 2023] and DeepChorus [He et al., 2022]. This dynamic approach provides a more precise and contextualized analysis of the music's emotional characteristics in different parts of the composition, allowing for a deeper understanding of emotional nuances over time.

Building on these established methods, this work aims to contribute to the MEVD field by:

- this work builds upon and extends the previous research conducted Panda and Paiva on static MER, a former MSc thesis student of the Music Emotion Recognition - Next Generation (MERGE) team;
- improve the emotion recognition process, by incorporating segmentation tools.

Table 1.1 summarizes the objectives ranked on a high, medium, and low priority scale, representing their importance in this work. All possible efforts are made to accomplish them.

Objective	**Priority**
Replication of previous work	High
Implementation and comparison of static and dynamic window sizes in classical approaches	High
Implementation and comparison of static and dynamic window sizes in DL approaches	High
Implementation and comparison of static and dynamic window sizes in Hybrid network approaches	Medium
MEVD - Review and experimentation with Cal500Exp	Low

Table 1.1: Objectives of this work.

To fully elucidate the scope of this undertaking, accomplishing the project's aims should address the subsequent research questions:

- Are the results presented in previously conducted work replicable?
- How do classical and DL approaches affect the MEVD performance?
- How do fixed and variable window sizes affect the performance in classical approaches?
- How do fixed and variable window sizes affect the performance in DL approaches?

1.3 Results, Contributions and Limitations

This section presents the key outcomes of the research, highlighting the significant contributions made and addressing the limitations encountered during the course of this work.

The research yielded several noteworthy results:

- Achieving **53.17%** F1-score in MEVD dataset using an Support Vector Machine (SVM) with the segments produced by the All-in-One tool;
- Achieving **43.10%** F1-score in MEVD dataset using an Convolutional Neural Network (CNN) with the segments produced by the All-in-One tool.

As for the main contributions made with this work:

- Exploring a structural segmentation tool: All-in-One;
- Extensive experiments in Classic ML and DL approaches were conducted using the All-in-One tool.

As for limitations found during this work:

- One major challenge is that although the All-in-One tool holds significant promise, it still requires further refinement to realize its full potential. It achieved a weighted average F-measure of 70.10% for segmentation accuracy on the MEVD dataset, demonstrating that structural segmentation can be imprecise.

The main difficulties found during this work:

- The volatile nature of the server where the experiments were conducted;
- Problems with replication of virtual environments with outdated libraries;
- The limited size of the database, which did not allow for fully leveraging deep learning approaches.

1.4 Organization, Planning and Resources

This section outlines the experimental setting of this research, including a summary of the tasks planned, the time allotted for each, and a review of the project's actual progression.

1.4.1 Experimental Environment

The experiments described were primarily carried out on a server shared with the team. Given the complex requirements of the deep learning models tested, Graphics Processing Unit (GPU) were necessary for their effective development and evaluation within a feasible timeframe. The specifications of the server are:

- Intel Xeon Silver 4214 CPU @ 2.20 GHz x 48
- 7x NVIDIA RTX A5000 24GB
- 3x NVIDIA RTX A6000 48GB
- 700GB RAM

Additionally, as the server is shared among multiple students, not all resources are accessible at all times, leading to varying levels of server activity that can affect the speed of task completion.

The methodologies primarily utilized a Python 3.8.19 virtual environment to ensure replicability. Data manipulation was performed using libraries, including Numpy and Pandas. Keras, Tensorflow, and PyTorch were employed to build and train DL models. Scikit-learn's implementations of ML algorithms were used, in addition to calculating necessary metrics to evaluate the methodologies' performance.

1.4.2 Organization

The appendix presents Gantt charts to contrast the planned timeline with the actual effort needed to complete each task. The charts are accompanied by a discussion of the reasons for any changes.

First Semester

The decision was made to dedicate the initial two months of the semester primarily to conducting a literature review to develop a thorough state-of-the-art. Additionally, this period will be used to gather the essential information required to achieve the project's goals. The rest of the first semester would be focused on getting to know the available MER datasets, replicating the classic approach work done by Panda and Paiva, and experimenting with and becoming familiar with segmentation tools like All-in-One.

During the literature review period, there was an opportunity to familiarize myself with the server that would host the experiments and to prepare the necessary setup for conducting them.

The focus of the whole semester was to develop a good foundation on the various approaches available to solve MER. Replication of previous work was started,

ensuring a foundation of the main tools for conducting all future experiments in the next semester.

Second Semester

For this semester, most of the tasks were already established with respective estimations of the time they would take to complete. These estimations, as can be seen by the final Gantt chart, were severely underestimated due to various factors.

The project started with the continuation of the traditional window-based MEVD approach, which was initially expected to take four weeks. However, due to my initial lack of expertise, it took an additional two weeks to complete. A similar delay happened with the window-based MEVD DL approach, as I needed extra time to become proficient with DL algorithms, extending the timeline by two more weeks.

After starting the structural-based MEVD classical approach, everything was going according to plan until an unplanned server update caused the WELMO server to shut down, corrupting all data as a result. All the previous experimental data was lost, and we now need to redo the experiments to recover the lost results. This setback has caused a six week delay in the structural-based MEVD approach. To prevent future data loss, we have implemented regular backups.

The structural-based MEVD DL approach was making progress when another server update corrupted the virtual environment I was working in. Much time was spent trying to restore it, which unfortunately led to significant delays and limited the time for new experiments.

Due to the remaining time and the considerable delays caused by the virtual environment issues, the hybrid approach was initiated much later than expected, and experiments with the Cal500Exp dataset could not be conducted.

1.5 Outline

This document structure follows:

Chapter 2 overviews fundamental concepts for understanding MER and MEVD. It begins with discussing what emotions are, the different types of emotions, and various emotion models (Section 2.1). The chapter then introduces machine learning, including its types and the specific algorithms used in this work (Section 2.2). It continues with an overview of the feature engineering process, highlighting key audio features and frameworks (Section 2.3). The chapter then discusses deep learning, focusing on relevant models applied in this research (Section 2.4). It then explains the concept of data augmentation, its purpose, and methods (Section 2.5). The chapter concludes with a discussion of the evaluation metrics used to assess model performance (Section 2.6) and a summary of the entire Section 2 (Section 2.7).

Chapter 3 provides an overview of the current state of research in MER and MEVD. It begins with a discussion of the various datasets used in the field, highlighting their importance and associated challenges (Section 3.1). The chapter then reviews the classical and deep learning approaches used for MER, outlining key methods and advancements (Section 3.2). Also, it discusses the work done in MEVD, exploring both classical and deep learning approaches and their applications (Section 3.3). Finally the chapter concludes with a discussion on segmentation tools (Section 3.4) and a summary of the entire Section 3 (Section 3.5).

Chapter 4 describes the methods and experiments conducted in this research. It begins by replicating previous work to establish a baseline (Section 4.1) and proceeds with experiments using segmentation tools to test their effectiveness in emotion detection (Section 4.2). The chapter outlines classical machine learning approaches (Section 4.3) and deep learning approaches (Section 4.4), before introducing a hybrid approach (Section 4.5). It ends with a summary of the entire Section 4 (Section 4.6).

Finally, chapter 5 summarizes the main conclusions drawn from this research and discusses potential directions for future work. It includes recommendations for improving MER and MEVD methodologies and identifies areas for further investigation to advance the field.

Chapter 2

Background Concepts

This chapter will delve into the fundamental concepts that form the backbone of our study of MER and MEVD. By exploring these key concepts, we can build a strong understanding of the principles and ideas that underlie our field of study.

2.1 Emotion

Emotions are a fascinating and complex topic in psychology. The term "emotion" has been used in various ways throughout history, leading to a lack of agreement on its definition and characteristics [Dixon, 2012].

In the upcoming sections, this work will explore emotions in-depth and provide a detailed overview of the various types of emotions and classification models used to identify them. The goal is to provide a comprehensive understanding of this intricate subject matter.

2.1.1 Definition of Emotion

The American Psychological Association proposes an adapted definition from Merriam-Webster for emotion, visually illustrated in Figure 2.1.

"Emotions are conscious mental reactions (such as anger or fear) subjectively experienced as strong feelings usually directed toward a specific object and typically accompanied by physiological and behavioural changes in the body." [Merriam-Webster, 2023].

2.1.2 Types of Emotion

"The distinction between emotion perception and emotion induction is important since it is possible to perceive emotional expression in music without necessarily being affected oneself." [Gabrielsson, 2001].

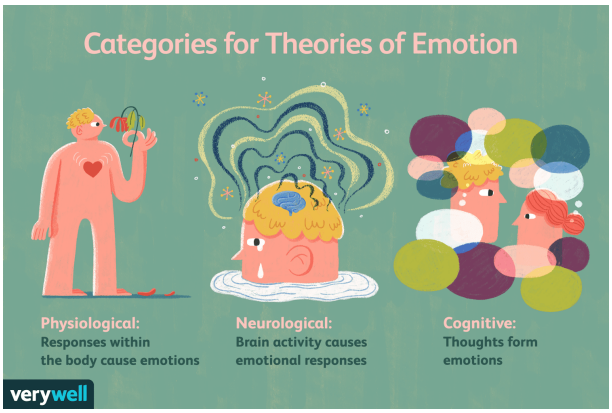


Figure 2.1: Theories of Emotion.
`https://www.verywellmind.com/theories-of-emotion-2795717`

Therefore, in MER, emotions are expressed, perceived, and felt. Expressed emotions refer to the emotions the author is trying to convey in his work. Perceived emotions in music refer to the emotion that a listener "observes" in a musical piece, which may or may not be different from the emotions that a musician intends to convey in a song from the emotions the listener feels. Finally, felt emotions in music refer to the emotional experience that the listener personally and subjectively feels while listening to a particular piece of music.

While expressed emotions in music are often accurately perceived by listeners, induced emotions can vary greatly depending on an individual's unique characteristics and personality. As a result, these emotions may differ between listeners and even within the same individual under different circumstances. In this respect, "the paradox of negative emotion refers to the phenomenon where music described in negative emotional terms, such as sadness or grief, is often judged as enjoyable." [Pannese et al., 2016].

As induced emotions are more subjective than perceived emotions, studies dealing with different kinds of emotions may create poor datasets. To address this issue, researchers primarily focus on perceived emotions, which have higher levels of agreement among listeners. However, subjectivity still exists, and measures must be taken to minimize it during the dataset creation process.

2.1.3 Emotion Models

Due to the ambiguous nature of emotions, a generic reference is necessary when categorizing music pieces. Music psychology has researched this topic, mainly focusing on the most suitable emotion taxonomy for modeling the emotional spectrum. There are two types of emotion models: categorical and dimensional. Categorical models classify emotions into specific categories: happiness, sadness, anger, and fear. On the other hand, dimensional models represent emotions through continuous dimensions such as valence and arousal, as discussed below.

Categorical/Discrete models

These models rely on [Ekman, 1992] concept of basic emotions, which posits that emotions are discrete and can be categorized. However, some researchers have challenged this theory by proposing alternative sets of emotions.

Hevner's Adjective Circle

Hevner's emotion model, initially developed by [Hevner, 1936], consists of 67 adjectives grouped into eight clusters that describe related emotional states. In the original model, each cluster contained a fixed number of adjectives. However, in later revisions made by other researchers, such as [Farnsworth, 1954] and [Schubert, 2003], the number of adjectives in each cluster was adjusted to include six to eleven adjectives per cluster. These revisions aimed to update the model by adding new terms and reorganizing the clusters to make it applicable to a broader range of music genres beyond classical music, addressing the limitations of Hevner's original genre-specific selection of adjectives.

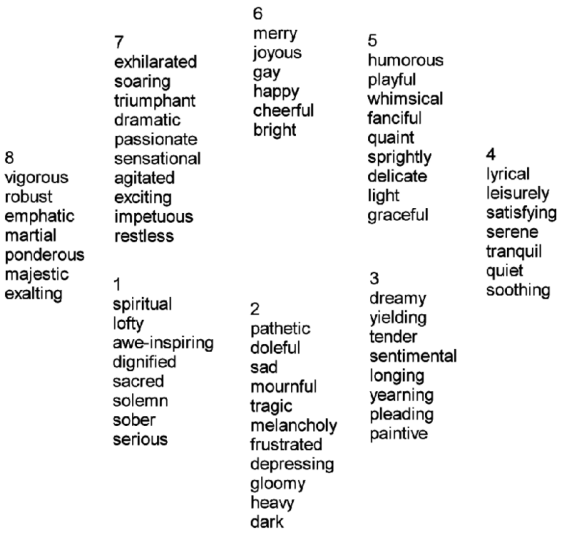


Figure 2.2: Hevner's Adjective Circle.
[Hevner, 1936]

GEMS

The Geneva Emotional Musical Scale (GEMS) study [Zentner et al., 2008] involved a series of experiments in which participants rated their emotional responses to different musical excerpts using a list of emotion terms. The researchers used this analysis to develop a categorical model of nine emotion categories consistently evoked by the music samples. These categories were Wonder, Transcendence, Nostalgia, Tenderness, Peacefulness, Power, Joy, Sadness, and Tension. The study's findings have significant implications for music emotion classification and understanding the emotional impact of music on listeners. Some have questioned the study's representativeness, as it selected only five genres to capture the entire spectrum of music, which may not fully reflect the diversity of musical expression.

Dimensional models

The emotions are represented in a multi-dimensional space, often with two dimensions, allowing for similarity judgments based on the distance between audio clips. The Russell Circumplex Model of Emotion [Russell, 1980] is widely recognized and influential in music emotion recognition.

Russell's Circumplex Model

The Circumplex Model of Affect [Russell, 1980] is a model that categorizes emotions based on their valence and arousal levels. Valence pertains to the polarity of emotion in terms of positive and negative states (also known as pleasantness), while arousal (also known as activity, energy, or stimulation level) refers to the activation or deactivation associated with an emotion. Russell even claimed that valence and arousal are the "core processes" of affect, constituting the raw material or primitive of emotional experience [Russell, 1980]."

The model proposes that emotions are within a circular structure, with adjacent emotions sharing similar valence and arousal levels, divided into four quadrants, each representing a different combination of valence and arousal levels: High Arousal, Positive Valence - excitement, enthusiasm, and ecstasy; High Arousal, Negative Valence - anger, fear, and anxiety; Low Arousal, Negative Valence - sadness, boredom, and depression; Low Arousal, Positive Valence - relaxation, contentment, and serenity.

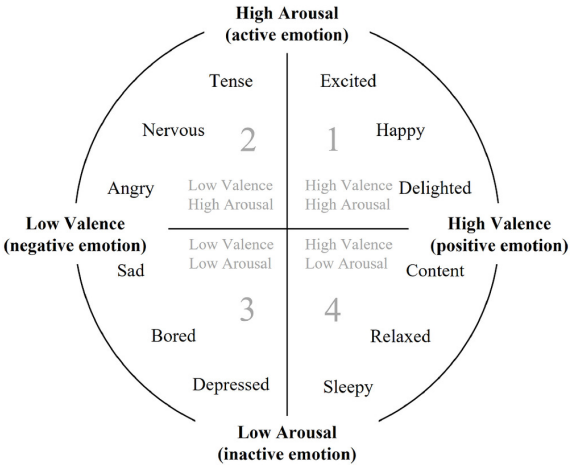


Figure 2.3: Russell's Circumplex Model.

Figure 2.3 shows that the intersection of arousal and valence dimensions forms a circular model with four quadrants, each representing a unique combination of valence and activation. Quadrant 1 represents positive, high-energy emotions like happiness. Quadrant 2 represents negative, high-energy emotions like anger. Quadrant 3 represents negative, low-energy emotions like sadness. Quadrant 4 represents positive, low-energy emotions like relaxation. This model suggests that emotions are interrelated and provides a framework for understanding and studying human emotions.

2.2 Machine Learning

ML is a subfield of Artificial Intelligence that aims to make computer systems learn and adapt from experience, like humans, creating algorithms that recognise data patterns and enabling systems to make informed decisions or predictions without explicit programming [Mitchell, 1997].

The ML process begins with data ingestion and preparation. This involves collecting data relevant to the problem, cleaning it to remove any errors or inconsistencies, and transforming it into a format suitable for the learning algorithm.

A suitable model, a mathematical representation of a real-world process, is selected once the data is pre-processed. The learning algorithm can adjust the model's parameters to improve performance, optimising these parameters based on a cost or loss function. This function quantifies the difference between the model's predictions and the actual data, providing a metric the algorithm seeks to minimise.

After the model training and satisfactory performance, the final step is deploying the model. This involves integrating the model into the existing production environment, where it can provide predictions on new data.

All of these steps are illustrated in Figure 2.4.

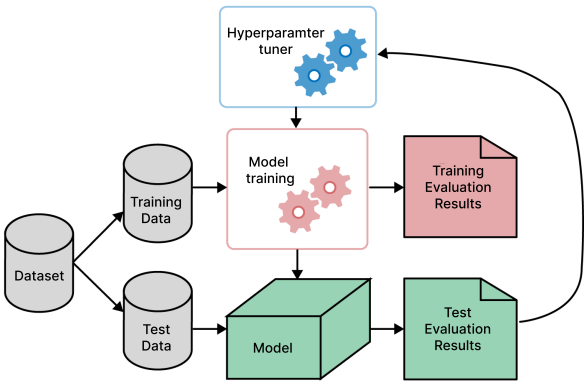


Figure 2.4: ML pipeline (adaptation).
[Omer et al., 2023]

Previous studies have explored the use of SVM ML models in music emotion recognition. One such example is the work of Panda and Paiva (2011), who developed a method for automatic emotion tracking in audio music through supervised learning and classification. Their approach predicted the quadrants of Russell's taxonomy and arousal and valence values of short segments in full songs, thus providing insights into the changes in emotion over time.

2.2.1 Machine Learning Paradigms

In ML, three primary categories exist: Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

Supervised Learning is a type of ML where the model is trained on a dataset that includes input variables and the corresponding correct outputs. It's analogous to a teacher-student scenario, where the model learns from labeled examples provided in the training data. The learning algorithm iteratively makes predictions and is corrected based on the actual outputs, leading to adjustments in the model's parameters. The goal is to optimize these parameters to minimize the discrepancy between the predicted and actual outputs, typically measured by a loss function. Supervised Learning can be further divided into two categories: regression, where the output is continuous, and classification, where the output is categorical [Bishop, 2006].

Unsupervised Learning, by contrast, involves training a model on a dataset without labels. The primary objective is to discover underlying structures or patterns within the data. This is achieved through methods such as clustering, where the algorithm groups similar data points, or dimensionality reduction, where the algorithm identifies the most essential features of the data. Unsupervised Learning is uncovering hidden labels or structures within the data [Bishop, 2006].

Reinforcement Learning is a different type of ML where an agent learns to make decisions by interacting with its environment. The agent takes actions based on its current state and receives rewards or penalties as feedback. The agent's objective is to learn a policy, a set of rules that guide its actions to maximize the cumulative reward over time. This learning process involves balancing exploration, where the agent tries different actions to gather information, and exploitation, where the agent uses the acquired data to make the best decisions [Sutton and Barto, 2018].

2.2.2 Machine Learning Algorithms

As this research focuses on a classification task to categorize different segments into four Russell's quadrants, this study will concentrate on supervised algorithms, namely SVMs.

Given its prevalence in most MER and MEVD studies [Panda and Paiva, 2011; Panda et al., 2020b], the experiments conducted in this work used SVMs, a robust set of supervised ML algorithms that can be used for various tasks such as classification, regression, and outlier detection.

As represented in Figure 2.5, the SVM model aims to find the most effective boundary between data classes. Its approach is to identify a hyperplane that offers the maximum possible separation between the two data classes. SVMs transform data into a high-dimensional feature space, establishing a linear decision boundary. Subsequently, SVMs identify the hyperplane that offers the most significant margin between the two data classes. The margin represents the dis-

tance between the hyperplane and the nearest data points from each class. Figure 2.5 represents this model.

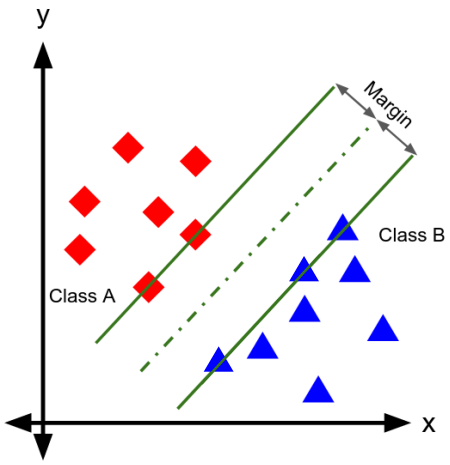


Figure 2.5: SVM Model Struture
`https://medium.com/it-paragon/support-vector-machine-regression-cf65348b6345`

SVMs aim to find an optimal hyperplane that maximally separates different classes in the feature space. The SVM algorithm finds the hyperplane by solving a quadratic optimization problem that seeks to minimize the norm of the weight vector subject to some constraints. These constraints ensure that each data point is on the correct side of the hyperplane.

2.3 Feature Engineering

Feature engineering is a crucial task within MER. It involves transforming raw data into meaningful features that improve the performance of predictive models. This process is essential for boosting the accuracy of ML models by effectively representing the underlying data. In MER, there is a significant focus on creating emotionally relevant features from audio and tasks such as feature scaling, integration, ranking, selection, and projection. This chapter will mainly concentrate on feature engineering for audio, as only audio features were used in this research.

2.3.1 Feature Engineering Overview

Feature engineering involves several steps to improve data representation for ML models. Key processes include:

Feature Extraction

ML algorithms rely on numerical data. However, inputs like audio files need to be converted into numerical values through feature extraction. Specifically for audio, it's essential to understand that sound is a series of waves representing air pressure changes over time. These waves have various properties, such as amplitude, frequency, and waveform complexity. Tools like the Fast Fourier Transform (FFT) can analyze these waves and develop numerical features like beats per minute or dominant frequency to convert the analog sound into digital form. These features are crucial for ML models to identify patterns and make accurate predictions.

Feature Scaling

Adjusting the range of feature values is typically done through normalization or standardization. Normalization transforms features to a common scale without distorting differences in the ranges of values. Standardization rescales data to have a mean of zero and a standard deviation of one.

Integration

Integration combines multiple features into common statistics. This is often done using metrics such as mean and standard deviation. Integration helps in summarizing the data more compactly, making it easier to analyze and use in ML models.

Ranking and Selection

Ranking and selection involve evaluating the importance of features and eliminating irrelevant or redundant ones. Techniques such as correlation analysis, mutual information, and feature importance scores from models like random forests or gradient boosting are commonly used.

Projection

Projection transforms feature spaces using techniques like Principal Component Analysis (PCA) to reduce dimensionality. PCA and similar methods help in identifying the most significant components of the data, thus simplifying the model and reducing overfitting.

2.3.2 Audio Features

Audio features are essential for understanding and analyzing the emotional content of music. These features capture various sound aspects that can be quantified

to predict a piece's emotional impact. Musical attributes can be grouped into four to eight different categories depending on the author (e.g., [Owen, 2000], [Meyer, 1973]), each representing a core concept. Here, we follow an eight-category organization, which includes melody, harmony, rhythm, dynamics, tone colour, expressive techniques, musical texture, and musical form.

- **Melody**: Melody is the sequence of pitches that form the main tune or theme of a piece, often the most memorable part of the music. It is characterized by the pitch, direction, and contour of the notes.
- **Harmony**: Harmony refers to the combination of different notes played simultaneously to create chords. It adds depth and supports the melody, creating feelings of tension or resolution.
- **Rhythm**: Rhythm refers to the pattern of sounds and silences in music, defining the timing and duration of notes. It sets the pace and flow, influencing how lively or calm a piece feels.
- **Dynamics**: Dynamics involve the variation in loudness within a piece of music, adding emotional intensity and contributing to the dramatic effect by ranging from very soft to very loud.
- **Tone Colour (Timbre)**: Tone colour, or timbre, describes the unique quality of a sound that distinguishes different instruments or voices. It is what makes a piano sound different from a violin, even when they play the same note.
- **Expressive Techniques**: These are methods used by performers to shape the music, affecting the transition and continuity between notes. Techniques such as staccato (short and detached) and legato (smooth and connected) impact the expressiveness and phrasing of the music.
- **Musical Texture**: Musical texture pertains to how different musical lines or layers are combined within a piece. It can range from simple (a single melody) to complex (multiple overlapping melodies), affecting the richness and complexity of the music.
- **Musical Form**: Musical form is the overall structure or organization of a piece of music. It determines how the music is arranged into sections and provides coherence and shape to the composition.

Standard Features

Standard features are foundational audio features commonly used in MER to capture basic characteristics of the audio signal. Examples of standard features across different musical dimensions include:

- **Melody**:
  - *Pitch Estimation*: Tracks the sequence of pitch values, capturing the melodic contour of the music.
  - *Predominant Pitch*: Extracts the Fundamental Frequency (F0) of the main melody line, the key emotional element.
- **Harmony**:
  - *Chromagram*: Maps energy distribution across the twelve pitch classes, analyzing harmonic content and key.
  - *Modality Estimation*: Determines whether the music is in a major (happy) or minor (sad) mode.
- **Rhythm**:
  - *Tempo Change*: Measures variations in the speed of a piece over time, influencing the music's emotional energy.
  - *Beats Loudness*: Estimates loudness at specific beats to understand rhythmic emphasis and intensity.
- **Dynamics**:
  - *Low Energy Rates*: Indicates the percentage of frames with less energy than the average, identifying softer, less intense sections.
  - *Loudness*: Represents the perceived intensity of sound, conveying the power and emotional weight of the music.
- **Tone Colour (Timbre)**:
  - *Zero Crossing Rate (ZCR)*: Measures the rate of sign changes in the waveform, indicating noisiness or texture.
  - *Mel-frequency cepstrum coefficients (MFCC)*: Analyzes spectral shape to capture the timbral qualities of sound.
- **Expressive Techniques**:
  - *Average Silence Ratio*: Measures the proportion of silence between notes, which can be used as an assessment of articulation, indicating how detached or connected the notes are.
- **Musical Form**:
  - *Similarity Matrix*: Assesses structural similarity between frames, identifying repeating sections and variations.
  - *Novelty Curve*: Highlights significant structural changes, marking transitions and new themes.

Novel Features

Novel features, as introduced by Panda et al., were developed to address the limitations of standard audio features, which are often low-level and derived directly from the audio waveform or spectrum. In contrast, humans naturally rely on higher-level musical concepts such as melodic lines, notes, intervals, and scores to perceive emotions in music. These novel features capture information about higher-level musical concepts such as melody, articulation and texture by explicitly determining musical notes, frequency, and intensity contours. This approach provides a more comprehensive understanding of the emotional content of music, bridging the gap between low-level audio descriptors and the listener's emotional perception.

- **Melody**:
  - *Register Distribution*: Analyzes how melody notes are spread across pitch ranges (e.g., soprano, bass), impacting emotional tone.
  - *Note Smoothness Statistics*: Indicates how close consecutive notes are in pitch, reflecting melody smoothness and emotional flow
- **Rhythm**:
  - *Note Duration Statistics/Distribution/Transition Ratios*: These features measure note duration ratios (short, equal, long) across the whole piece and per second, providing insights into rhythmic complexity.
- **Dynamics**:
  - *Ratios of Note Intensity Transitions*: Measures transitions between note intensities: higher, lower, or equal, to capture dynamic changes.
  - *Crescendo and Decrescendo Metrics*: Based on the intensity difference between note halves, these metrics count crescendo/decrescendo notes and sequences, detailing intensity changes over time.
- **Expressive Techniques**:
  - *Articulation Features*: Indicates how close consecutive notes are in pitch, reflecting melody smoothness and emotional flow.
- **Musical Texture**:
  - *Music Layers Statistics*: Estimates the number of simultaneous musical layers (F0s) in each frame, providing textural complexity.
  - *Ratio of Musical Layers Transitions*: Tracks transitions between different musical textures, showing how the texture evolves.

2.3.3 Audio Frameworks

Many of the standard audio features discussed in the previous section are implemented in various audio analysis frameworks developed over the years. This section provides a brief description of two notable frameworks, Marsyas and the MIR Toolbox, highlighting their strengths and weaknesses in extracting these features for analyzing the emotional content of music.

Music Analysis Retrieval and Synthesis for Audio Signals (Marsyas) is an open-source framework [Tzanetakis and Cook, 2000]. It offers integration with graphical interfaces, acoustical and statistical feature extraction, and classifier training, making it a versatile tool for music analysis. However, according to a study [Panda and Paiva, 2011], notable disadvantages such as a lack of comprehensive documentation, a complex API, and syntax difficulties can make it challenging for users to utilize the framework effectively.

The MIR Toolbox is a MATLAB-based framework [Lartillot et al., 2008] that offers a wide range of algorithms for music feature extraction. It is highly flexible, allowing users to combine various feature extraction modules to create custom analyses. The toolbox supports low-level and high-level audio feature extraction and is well-documented, with visualization tools and the ability to perform bulk extractions from multiple audio files. However, its reliance on MATLAB and the Signal Processing Toolbox makes it resource-intensive, which can limit its practicality for real-time feature extraction applications.

As the limitations of traditional frameworks become more apparent, mainly when dealing with large, complex datasets, there is an increasing need for more automated and scalable solutions. This is where DL techniques come into play. By leveraging the power of the Neural Network (NN), DL models can automatically learn and extract meaningful features from raw audio data, reducing the reliance on manual feature engineering and enabling more robust analysis of music's emotional content. The following section explores how DL is transforming the field of music emotion recognition.

2.4 Deep Learning

DL, a field of ML, uses deep NNs to learn from large, high-dimensional data. These networks, structured in layers, transform input data into meaningful output, enabling the model to learn complex data representations. Unlike traditional ML, DL can automatically discover the most relevant features from the data, removing the need for researchers to extract features manually.

However, the effectiveness of DL models is constrained by the necessity of large, high-quality datasets and substantial computational resources for training. This is a significant limitation, especially in MER fields. In MER, large annotated datasets are rare due to the subjective nature of emotional labels and the difficulty in gathering extensive labeled data. Additionally, resource constraints, including the need for powerful hardware and long training times, further exacerbate the

challenges.

In the upcoming sections, we will delve deeper into the intricacies of DL, exploring different DL models.

2.4.1 Artificial Neural Networks

NNs are computational models inspired by the structure and functioning of the human brain, designed to learn and make predictions from data. They consist of interconnected nodes, called neurons, organized into layers.

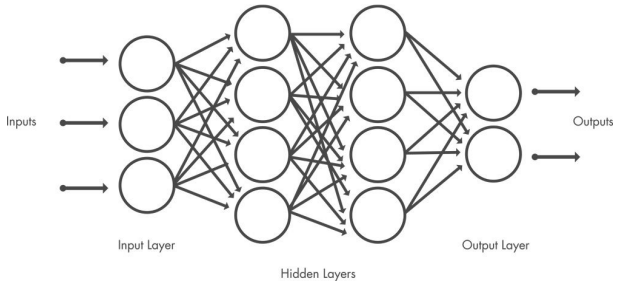


Figure 2.6: Example of a simple, fully-connected NN architecture. `https://www.mathworks.com/discovery/convolutional-neural-network.html`

As seen in Figure 2.6, the NN architecture consists of an input, hidden, and output layer.

The input layer is the first layer of a NN. Each neuron in this layer represents a feature of the input data.

The hidden layers, positioned between the input and output layers, play a crucial role in learning and extracting features from the input data through complex combinations of weighted connections. In a dense network, each neuron in the hidden layers receives inputs from all neurons in the preceding layer, applying a set of weights and biases to them. In contrast, a sparse network has more limited connections, where each neuron only receives inputs from a subset of neurons in the previous layer. After this process, the results are passed through a non-linear activation function, enabling the network to capture more complex patterns in the data.

Weights and biases are learned parameters that determine the strength of connections between neurons in a NN. Activation functions, such as sigmoid, Rectified Linear Unit (ReLU), and softmax, introduce non-linearity to the output of neurons, allowing the network to model complex relationships between inputs and outputs. ReLU sets negative values to zero, addressing the vanishing gradient problem, accelerating training, and reducing computational complexity. Sigmoid outputs probabilities between 0 and 1, making it useful for binary classification tasks. Softmax normalizes a vector into a probability distribution, which is ideal for multi-class classification tasks.

The output layer is the final layer of the NN that generates predictions. The number of neurons depends on the problem. In binary classification, there may be one neuron for each class, while in multiclass classification, there may be multiple neurons. One-hot-encoding is applied to labels in order to compare the output of the network.

Training a NN involves several key methods. During feedforward, input data traverses through the network layer by layer. Neurons compute weighted sums of inputs and apply activation functions to generate predictions.

BackPropagation (BP) fine-tunes NN weights by comparing predicted and actual outputs, guiding the Gradient Descent optimization algorithm during training, and adjusting weights iteratively to minimize loss function and improve accuracy. It ensures the NN learns from errors and moves towards optimal weights for accurate predictions. In addition to Gradient Descent, other optimization algorithms like Adaptive Moment Estimation (Adam) and Stochastic Gradient Descent (SGD) also play a critical role in training NNs.

The loss function evaluates model performance on the training data by measuring the discrepancy between predicted outputs and actual targets. Common loss functions include Mean Squared Error (MSE) and Cross-Entropy Loss.

2.4.2 Convolutional Neural Networks

CNNs are composed of multiple layers that can detect distinct features within data. These networks employ filters that apply to data at various resolutions, and each filtered data output serves as the next layer's input. The filters initially learn to identify essential features and then gradually progress towards more complex features that can precisely define the data [Goodfellow et al., 2016].

The input goes to a convolutional layer, where filters are applied to the data to extract features. These filters start by detecting simple features like edges and colours, and as the data passes through the network, they can see more complex features like shapes or patterns.

Once the convolution process is complete, an activation function is applied to each element of the feature map.

Then, the pooling layer downsamples the data, reducing its dimensionality and allowing it to draw assumptions about features contained in the sub-regions binned. This layer reduces the computational cost by significantly reducing the number of parameters.

After several convolutional and pooling layers, the high-level reasoning in the NN is done via fully connected layers. In a fully connected layer, neurons connect to all activations in the previous layer, as seen in regular NNs. They use these features to classify the input image into various classes based on the training dataset.

A CNN comprises several crucial components in addition to its layers. One of

these essential components is the flattening step, which transforms the feature map into a one-dimensional matrix. This matrix serves as the input for an appended Artificial Neural Network (ANN), which is responsible for generating the final prediction of the CNN.

An illustration of the pipeline for classification using a CNN can be seen in Figure 2.7

Finally, the network learns the optimal filters to apply during the convolution operation through BP.

CNNs were utilized in the research to learn features from songs and classify them into different emotional quadrants according to Russell's model.

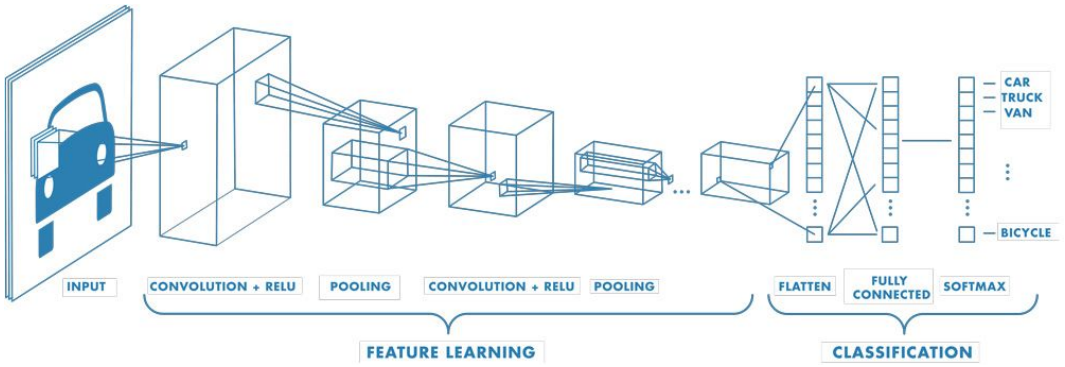


Figure 2.7: Visualization of CNN pipeline.
`https://www.mathworks.com/discovery/convolutional-neural-network.html`

2.4.3 Recurrent Neural Networks

Recurrent Neural Network (RNN) are a type of ANN designed to process sequential data by retaining memory of past inputs. Unlike traditional feedforward NNs, which process data sequentially without retaining memory, RNNs have connections between nodes that form directed cycles, allowing them to exhibit temporal dynamic behaviour.

There are four different types of RNNs, distinguished based on how far back the output of a neuron is passed within the network: direct-feedback-network, indirect-feedback-network, lateral-feedback-network, and complete-feedback-network [DatabaseCamp, n.d.]. Depicted in Figure 2.8 is an example of indirect-feedback-network architecture.

One significant advantage of RNNs is their ability to capture long-range dependencies in sequential data by propagating information through time. This capability allows RNNs to model context and temporal relationships effectively. However, RNNs face several challenges, including the vanishing and exploding gradient problems.

The vanishing gradient problem happens when gradients become very small while training and the network fails to learn long-range dependencies. On the

other hand, the exploding gradient problem occurs when gradients grow exponentially, causing instability during training and causing the weights to update erratically, making it difficult for the network to converge to an optimal solution.

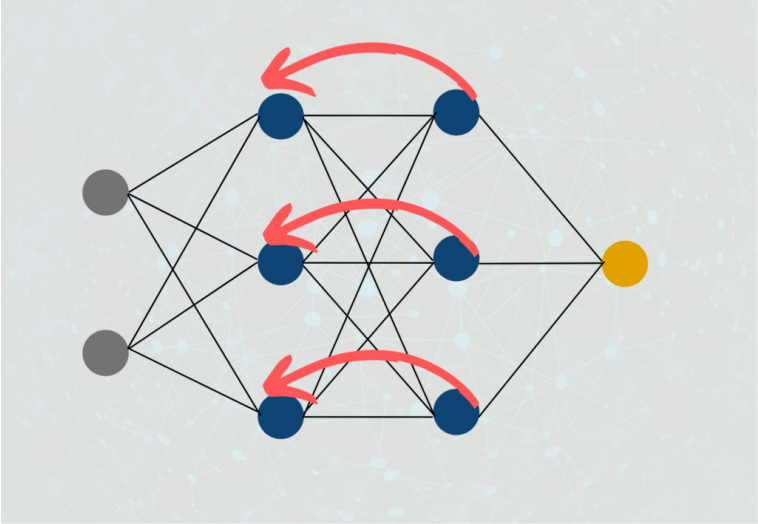


Figure 2.8: Visualization of a indirect-feedback-network (RNN).
`https://databasecamp.de/en/ml/recurrent-neural-network`

RNNs may also struggle with short-term memory, hindering their performance on tasks requiring modeling complex temporal patterns with long-range dependencies.

Long Short-Term Memory Networks

Long-Short Term Memory (LSTM) is a type of RNN capable of learning long-term dependencies in sequence prediction problems.

LSTM solves the vanishing gradient problem through its unique architecture. The input, forget, and output gates control the flow of information and gradients within the network, deciding what information should be kept, discarded, or passed on to the next time step. This selective memory feature helps to prevent the gradient from vanishing during BP.

The LSTM architecture consists of a cell and three types of gates: an input gate, an output gate, and a forget gate, as shown in Figure 2.9.

The cell is responsible for retaining values over arbitrary time intervals. The forget gate, which looks at the current input and the previous hidden state, decides what information should be discarded from the cell state by outputting a number between 0 and 1 for each number in the cell state. These numbers indicate the degree to which each piece of information should be retained (closer to 1) or forgotten (closer to 0).

The input gate updates the cell state with new information. It involves two parts: a Sigmoid layer called the “input gate layer” that decides which values will be

updated, and a tanh layer that creates new candidate values that could be added to the state. The output gate decides what the next hidden state should be. The hidden state, which contains information on previous inputs, is influenced by the output gate, which determines what information it should carry to the next part of the sequence.

While LSTM networks are robust for handling sequential data and learning long-term dependencies, their effectiveness still relies heavily on the availability of large, diverse datasets. Data augmentation techniques can be employed to artificially increase the variety of training data to enhance model performance and generalisation.

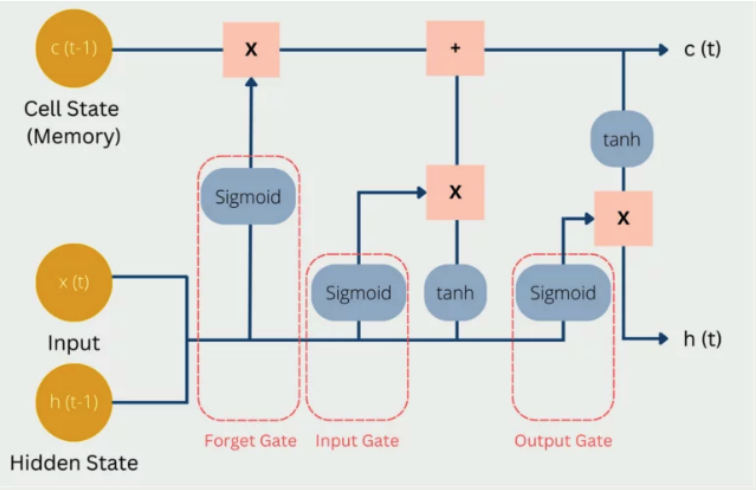


Figure 2.9: Visualization of LSTM architecture.
`https://databasecamp.de/en/ml/lstms`

2.5 Data Augmentation

Data augmentation is a technique used to increase the diversity and size of training datasets by applying transformations directly to the raw audio signal. These transformations help improve the robustness and generalization of DL models. Below are some commonly used audio augmentation classic methods:

- **Time Shifting:** Randomly shifts the audio sample in time, introducing silence at the beginning or end, making the model invariant to slight temporal misalignments.
- **Pitch Shifting:** Alters the pitch of the audio sample up or down by a random amount, such as a complete or half tone, to simulate variations in vocal pitch or instrumental tuning.
- **Time Stretching:** Changes the tempo of the audio without affecting the pitch, by randomly speeding up or slowing down the playback, helping the model adapt to variations in speed.

- **Power Shifting:** Adjusts the volume of the audio sample by increasing or decreasing the intensity by a certain number of Decibels (dB), making the model robust to changes in loudness.

These augmentation techniques are essential for training DL models to handle real-world variations in audio, improving their accuracy and reliability in tasks such as music emotion recognition.

While data augmentation techniques enhance the diversity and robustness of training data, understanding the structure of the audio is equally important for accurate analysis and model training. This is where segmentation tools come into play. By breaking down audio tracks into meaningful segments, these tools help capture music's emotional and structural nuances.

2.6 Evaluation Metrics

This section explores the various metrics crucial for evaluating hyperparameters during model training, assessing model performance post-training, and validating hypotheses. Also, it highlights the typical metrics used in the experimentation phase.

When creating a DL or ML model, it is crucial to select the correct hyperparameters. These include the number of epochs, batch size, optimizer, and learning rate for DL and cost, gamma, and kernel for ML. To evaluate these hyperparameters, we monitor the loss function, which measures the error between predicted and actual values and aims to minimize this error.

Optimization functions such as Bayesian search were employed to achieve this goal. Bayesian search is a hyperparameter optimization technique that uses past evaluation results to model the relationship between hyperparameters and model performance. Based on this model, it iteratively selects new hyperparameters, focusing on areas that are likely to improve performance, making it more efficient than random or grid search methods.

During the experimentation phase, the primary objective of optimization was to maximize the F1 Score. This metric considers both precision and recall, making it a good indicator of the model's overall accuracy.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision measures the proportion of true positives among all the predicted positive examples, while Recall measures the proportion of true positives among all the actual positive examples.

For example, our model predicts that 80 images are dogs, out of which 75 are actually dogs (true positives) and 5 are cats (false positives). Precision in this case would be calculated as

$$\text{Precision} = \frac{75}{75+5} = \frac{75}{80} = 0.9375$$

This means that our model is correct in predicting an image being a dog around 93.75% of the time.

In another scenario, suppose there are 100 actual dog images in the dataset. Our model identifies 75 as dogs (true positives) and misses 25 (false negatives). Recall is calculated as:

$$\text{Recall} = \frac{75}{75+25} = \frac{75}{100} = 0.75$$

This indicates that our model correctly identifies approximately 75% of all actual dog images in the dataset.

It is possible to analyze the performance of each individual class by using the Confusion Matrix of the predicted and actual targets, along with the F1 Score for each class and the overall F1 Score, as seen in 2.10.

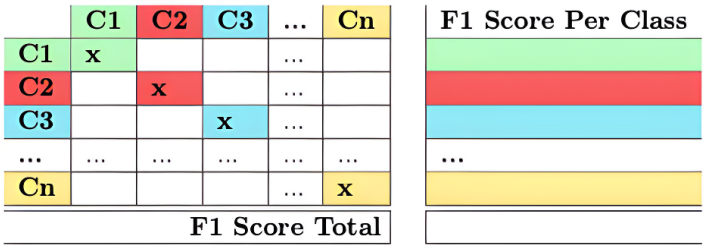


Figure 2.10: Generic Confusion Matrix with F1 Score example. [Louro, 2022]

For this study, recall, precision, and F1-score were selected because they provide a balanced view of the model's performance across all quadrants. Recall is particularly useful for understanding the model's ability to correctly identify all relevant instances within a quadrant, while precision measures how accurate the model's positive predictions are. The F1-score, as a combination of these two metrics, offers a single, balanced metric to assess classification performance across various quadrants.

Additionally, the confusion matrix was computed to visualize the distribution of classification errors across the four quadrants. This matrix allows us to see how many clips were misclassified among the quadrants, offering more profound insights into the model's performance at a granular level. By analyzing misclas-

sifications, we can better understand specific areas where the model struggles, whether with certain quadrants or types of music.

2.7 Summary

Chapter 2 provides a comprehensive overview of fundamental concepts for understanding MER and MEVD. The chapter begins by defining emotions and discussing their relevance to music, explaining the distinction between perceived and induced emotions, and highlighting how music can convey emotions intended by the artist and evoke different emotional responses in listeners.

The chapter then explores various models for categorizing emotions to provide a structured understanding. It covers categorical models like Hevner's Adjective Circle, which organizes emotions into discrete categories, and dimensional models like Russell's Circumplex Model, which represents emotions on a continuum using dimensions such as valence and arousal. These models offer valuable frameworks for understanding how emotions are represented and perceived in musical contexts.

Following this, the chapter introduces the fundamentals of ML, which play a crucial role in MER. It describes various approaches, including supervised, unsupervised, and reinforcement learning, focusing on supervised learning due to its ability to leverage labeled datasets for precise emotion classification tasks. This introduction to ML sets the stage for understanding how these techniques can be applied to analyze and interpret emotional content in music.

Transitioning from the basics of ML, the chapter emphasizes the importance of feature engineering in MER. Feature engineering involves processes such as feature scaling, integration, ranking, selection, and projection, which transform raw audio data into meaningful features that enhance the accuracy of predictive models. The chapter provides an in-depth look at both standard and novel audio features across various musical attributes, including rhythm, dynamics, expressive techniques, melody, harmony, tone colour (timbre), musical texture, and musical form. These features are critical for capturing the emotional nuances of music. Additionally, the chapter discusses audio frameworks like Marsyas and MIR Toolbox, which facilitate the extraction and analysis of these features, thereby supporting the feature engineering process.

Following feature engineering, the chapter discusses DL, highlighting its ability to automatically learn and extract relevant features from raw data. It delves into specific DL models such as CNNs and RNNs, including LSTM networks, explaining their roles in analyzing and interpreting audio data.

To further enhance the effectiveness of DL models, the chapter highlights the role of data augmentation. By applying transformations such as time shifting, pitch shifting, time stretching, and power shifting directly to raw audio signals, data augmentation artificially expands the diversity of training datasets. This process makes models more robust and better equipped to handle variations in

real-world audio data, ultimately improving their generalization capabilities.

Finally, it emphasizes the importance of evaluation metrics such as precision, recall, and F1-score in assessing the performance of MER models, noting that these metrics provide a balanced view of the model's accuracy and ability to correctly classify emotional segments within music.

Chapter 3 will further develop the foundational knowledge from chapter 2 by examining the latest advancements in MER models. We will explore different approaches and methodologies used in the field and provide a detailed review and constructive evaluation of existing research. This transition will help us understand the ongoing challenges and latest improvements in MER and MEVD, allowing for a deeper understanding of how these technologies can be applied and enhanced in future studies.

Chapter 3

State of the Art

This chapter aims to provide readers with an overview of the fundamental concepts of MER models while examining the diverse range of approaches to these models. Furthermore, a detailed review and constructive evaluation of the existing research in this field will be presented.

3.1 MER Datasets

Datasets are essential in advancing research in MER and MEVD by providing a foundation for training and evaluating ML models. In static MER, the goal is to identify a song's predominant emotion by analyzing a smaller excerpt that best represents the overall emotional content of the music. This approach allows for classifying a song's emotion based on a segment that captures the most dominant emotional expression throughout the piece. In contrast, MEVD focuses on understanding emotional variation over time in a complete song and thus requires the annotation of entire songs. Segmentation of songs is needed to gain a more nuanced understanding of emotion changes throughout the piece.

Numerous difficulties in current datasets used in music emotion recognition delay the growth of this field. Limited size and diversity are common issues that affect the applicability of models. The subjective nature of emotions, the inconsistency in annotations, and imbalances in class distribution are hurdles that can impact the quality and reliability of datasets. Furthermore, capturing temporal dynamics and ensuring privacy in datasets present ongoing complications. Copyright restrictions, lack of standardization, and domain specificity complicate the use of these datasets. Moreover, the continuous development of technology and research methodologies requires ongoing efforts to improve the quality and relevance of music emotion datasets.

Developing MER datasets poses significant challenges. The process includes a laborious manual annotation task associating emotions with each audio snippet or song segment, making it time-consuming. For MEVD, defining segmentation protocols introduces complexities related to song structure. Addressing the subjective nature of music-induced sentiments and inter-listener variability requires

employing multiple annotators, retaining only segments with high agreement.

Ensuring diversity in dataset characteristics, such as genre distribution and artists, is essential for representativeness. Taking proactive measures, such as deliberately choosing songs positioned away from central emotional tendencies on the Russell plane and rigorously validating annotations, helps alleviate the impact of emotion subjectivity. The dataset creation process typically begins with a large set of songs. Still, it often results in a more focused subset, emphasizing the importance of careful selection and refinement to ensure quality and relevance in the final dataset.

The subsequent sections will furnish descriptions of the primary datasets utilized in Static MER and MEVD, arranged according to their year of publication.

RWC

The Real World Computing (RWC) Music Database [Goto et al., 2002] is a copyright-cleared music database available to researchers as a common foundation for research.

The RWC Music Database contains 100 complete songs with manually labeled section boundaries.

The RWC Music Database, while valuable, has its limitations. It's relatively small, with around 100 songs, which may not suffice for research, especially deep learning models that need large datasets. The specific collection of tracks could introduce bias, potentially limiting model generalization. Lastly, it might lack diverse annotations beyond section boundaries, limiting its usefulness for specific research tasks.

Million Song Dataset

The Million Song Dataset (MSD) [Bertin-Mahieux et al., 2011] is an invaluable resource for MER, offering access to audio features and metadata for a vast collection of one million contemporary popular music tracks. This freely available dataset presents a comprehensive and detailed perspective to support various research tasks within the realm of MER, curated from The Echo Nest, a large music database acquired by Spotify shortly after the development of this dataset.

The dataset underwent rigorous data cleaning procedures, including disposal of duplicates, error correction, and imputation for missing values. This dataset offers significant value to researchers and enthusiasts in the field of MER due to its extensive metadata and audio analysis for one million legally available songs to The Echo Nest. It provides insights into contemporary popular music, including trends in genre, instrumentation, production techniques, and patterns in songwriting and performance styles.

However, it is worth acknowledging that the dataset has limitations, particularly in diversity, as it lacks representation of world, ethnic, and classical music. It is important to note that the dataset provides only features and metadata, with annotations based on uncontrolled, user-generated tags from the Last.fm music social network, which are often ambiguous and inconsistent. Despite this, its

richness in contemporary popular music positions it as an asset for advancing research and understanding within the MER domain.

MedleyDB dataset

MedleyDB [Bittner et al., 2014] is a meticulous, multitrack dataset designed for annotation-intensive research for MIR. The dataset contains 122 songs, with 108 featuring melody annotations, making it an ideal tool for various MIR applications, including instrument identification, source separation, and automatic mixing. MedleyDB was curated to address limitations in existing multitrack datasets, such as small size, lack of variety in genre, and non-uniform formatting.

The selection of the 122 songs was deliberate, focusing on diversity in genre, instrumentation, and recording quality. The multitrack recordings underwent thorough annotation processes, with melody F0 annotations, which refer to marking or labeling the F0 of the melody in a musical recording, corrected manually by annotators using a state-of-the-art extraction algorithm.

CAL500exp dataset

The new dataset introduced in this section is called CAL500exp [Wang et al., 2014], an expansion of the CAL500 dataset.

The CAL500exp dataset is unique because it uses variable-length segments, ranging from 3 to 16 seconds, while other segment-level datasets use fixed-length segments. It has 3,223 segments from 500 tracks, each annotated with 67 expert-defined tags covering eight semantic categories: emotion, genre, best-genre, instrument, instrument solo, vocal style, song characteristic, and usage.

Labels are obtained by the decision of each tag by "majority voting" over at least three paid university students. The dataset is available upon request to the authors of the dataset.

Bi-Modal dataset

The dataset [Malheiro et al., 2016], initially formed by merging a lyrics dataset of 180 samples and an audio dataset of 162 clips from diverse sources, resulted in a bimodal dataset containing 133 songs with both audio and lyrics. Thirty-nine annotators independently categorized the audio and lyric, assigning valence and arousal values utilizing a discrete Russell's Arousal/Valence (A/V) model.

Features were organized based on fundamental musical concepts, with audio features categorized under rhythm and melody, and lyric features derived from state-of-the-art methods like bag-of-words, part-of-speech tagging, and grammatical class occurrences.

The final dataset, categorized into four quadrants based on valence and arousal, raised concerns about potential imbalances due to its reduced size. Despite these limitations, the dataset proposed by Malheiro et al. stands out for its bimodal approach, combining audio and lyrics to enhance performance in emotion recognition tasks. Even with a smaller sample count, its versatility underscores its relevance for independent use in audio-based MER research.

DEAM dataset

The MediaEval Database for Emotional Analysis of Music (DEAM) dataset [Aljanaki et al., 2017] contains 1,802 music tracks and song excerpts from various Western popular music genres. Each song and song excerpt is annotated with valence and arousal scores, representing positive or negative emotions and energy levels in music, respectively. The song excerpts are 45 seconds, while the dataset includes 58 full-length songs.

Regardless, the DEAM dataset used in the benchmark is limited to Western popular music genres, with annotations based on perceived emotion, which reflects subjective human opinions. Consequently, the challenge with annotating music is that there can be a lack of consensus among annotators, as well as the presence of unclear audio clips. Additionally, since emotions in music can be perceived and interpreted differently across genres and cultures, the lack of detailed information on the distribution of annotators may contribute to variability in the annotations.

4QAED

The 4-Quadrant Audio Emotion Dataset (4QAED) dataset [Panda et al., 2018] was created by mapping annotations to the quadrants of Russel's model, using Warriner's list of adjectives, the clips and their corresponding annotations originated from the AllMusic API.

After filtering and discarding poor-quality clips, the balanced dataset contained 225 samples for each quadrant, totaling 900 samples. It is worth noting that this dataset suffers from issues such as its small size and lack of quality sources.

The core objective of the research, exemplified by 4QAED and other datasets, highlights the essential need to create more comprehensive and high-quality datasets for both MER and MEVD.

Harmonix

The Harmonix dataset [Nieto et al., 2019] is a comprehensive collection of annotations for over 900 tracks of Western popular music. The dataset includes annotations of beats, downbeats, and functional segmentation. It also contains additional metadata, such as MusicBrainz identifiers.

The dataset covers a wide range of Western popular music, with a strong emphasis on Pop, EDM, and Hip-Hop; the duration of these tracks varies as they are full-length songs.

It can be used for both static music emotion recognition and music emotion variation detection.

MERGE Audio Complete

The MERGE Audio Complete dataset [Louro et al., 2024b] consists of 3,554 30-second song excerpts distributed across Russell's quadrants: 875 songs in quadrant 1 (Q1), 915 songs in quadrant 2 (Q2), 808 songs in quadrant 3 (Q3), and 956 songs in quadrant 4 (Q4), as illustrated in Figure 3.1.

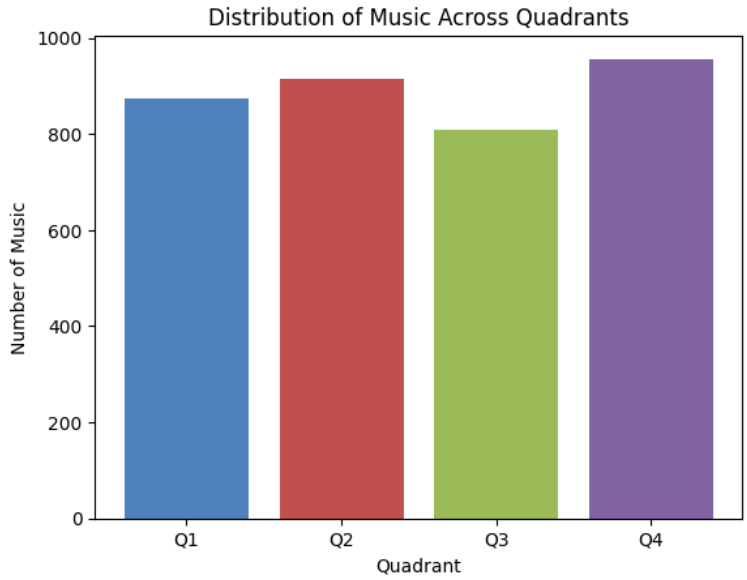


Figure 3.1: Complete dataset distribution across quadrants.

MERGE Audio Balanced

The MERGE Audio Balanced dataset [Louro et al., 2024b] provides an even distribution of samples across the four quadrants. It consists of 3,232 samples, with 808 songs in each quadrant (Q1, Q2, Q3, and Q4), as illustrated in Figure 3.2.

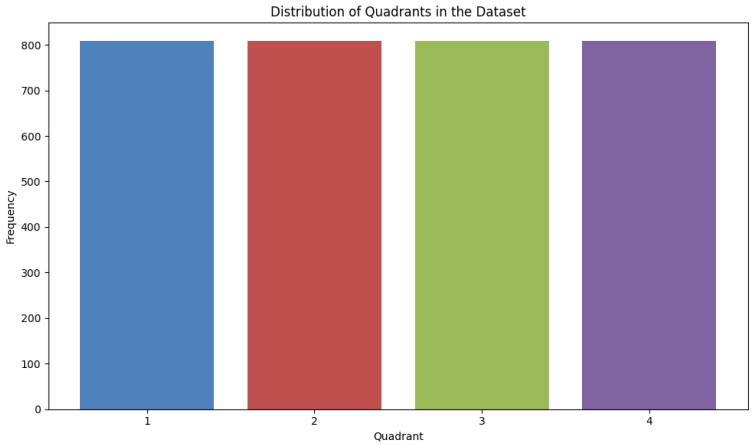


Figure 3.2: MERGE Audio Balanced dataset distribution across quadrants.

MEVD Panda

The MEVD Panda dataset consists of 29 songs distributed across Russell's quadrants: 10 songs in quadrant 1 (Q1), 7 songs in quadrant 2 (Q2), 2 songs in quadrant 3 (Q3), and 10 songs in quadrant 4 (Q4).

Name	Type	Emotion Taxonomy	Audio Duration	Size	Notes/Observations
RWC	Static MER	Emotional tags provided by users	full-length audio clips	100 audio clips	Limited size.
MSD	Static MER	Emotional tags provided by users	30 seconds audio clips	1000000 audio clips	Variety poses a challenge, with limited representation of global, cultural, and classical music.
MedleyDB dataset	Static MER	User-contributed emotional tags.	20 to 600 seconds audio	122 audio clips	Potential inaccuracies in annotations, the utilization of short excerpts rather than complete songs, and a restricted number of excerpts.
CAL500exp	MEVD	67 tags	3 to 16 seconds audio clips	500 songs divided into various clips.	The dataset remains relatively compact when compared to other music datasets, such as the MSD
Bi-Modal	Static MER	Russell's A/V Model	30 seconds audio clips	162 audio clips and 133 bi-modal	Despite well-executed annotations, the dataset size remains notably small, rendering experimental results less impactful.

(...continued from previous page)					
Name	**Type**	**Emotion Taxonomy**	**Audio Duration**	**Size**	**Notes/Observations**
DEAM	Static MER and MEVD	Russell's A/V Model	45 seconds clips	1744 audio clips and 58 full songs	Annotations rely on the subjective perspectives of human annotators and may not accurately capture the true emotional content of the music.
4QAED	Static MER	Russell's A/V Model	30 seconds audio clips	900 audio clips	Equally distributed among quadrants.
Harmonix	Static MER and MEVD	Russell's A/V Model	full-length audio clips	912 audio clips	Limited genre classification.
MERGE Audio	Static MER	Russell's A/V Model	30 seconds audio clips	A maximum of 3,554 audio clips	The limited dataset size restricted the full exploration of deep learning experiments.
MEVD Panda	MEVD	Russell's A/V Model	full-length audio clips	29 audio clips	Unevenly distributed across Russell's quadrants.

Table 3.1: Dataset's Review.

The common issues across most MER datasets include their limited size, which can hinder the training of robust models, especially deep learning ones. They may also contain biases due to the specific collection of tracks, potentially limiting model generalization. Furthermore, some datasets might lack diverse annotations, limiting their applicability for certain research tasks.

The dataset that stands out most is the 4QAED dataset for achieving the challenging task of obtaining equal distribution among the quadrants. The development of more comprehensive, high-quality, and equally distributed datasets would propel the field of MER forward, leading to more robust and accurate emotion recognition models.

3.2 Static MER

Static MER is a subset of MIR that focuses on determining and categorizing the dominant emotional content of a musical piece. This approach considers the entire musical composition as a single entity and assigns it a single emotional label or a set of labels.

This section offers a comprehensive overview of methods employed to address the challenge of emotion classification in Static MER.

3.2.1 Classical Approaches

To classify the emotional content of music, practitioners typically employ a recognition process that involves extracting various features such as tempo, pitch, timbre, and rhythm. Subsequently, these features become the basis for training ML algorithms, allowing for the classification of music's emotional content into categories, such as the four quadrants of the Russell model.

The accuracy hinges on factors like the quality of the training data, feature selection, and the complexity of the employed ML algorithm.

In the study "Popular Music Retrieval by Detecting Mood" by [Feng et al., 2003], the researchers utilized a dataset comprising 223 modern popular music pieces. Each piece is labelled with one of four emotion classes: happiness, sadness, anger, and fear. Extracting three audio features from these songs, they trained an ANN using a split dataset of 200 training songs and 23 testing songs.

The NN achieved commendable classification accuracies of 86% for happiness, 75% for sadness, and 83% for anger. However, the accuracy for the fear emotion class significantly lagged, reaching only 25%. This challenge, coupled with the constraints of small datasets and feature sets that might not comprehensively capture the diverse musical dimensions influencing emotion perception, elucidates this outcome.

[Meyers, 2007] developed a mood-based music classification and exploration system that uses audio files and lyrics to classify the emotional content in a song.

The method employs Russell's model and uses five features: mode, harmony, tempo, rhythm, and loudness. A decision tree algorithm is used for preliminary classification, followed by a K-nearest Neighbor (KNN) algorithm to classify the song into eight categories. The output of the KNN algorithm is combined with the affective value of the lyrics to predict the song's global emotion. The dataset used for this study was a private collection composed of 372 songs.

Although the study results were considered positive, a critical analysis reveals that the authors did not offer a statistical method to assess the performance of the models.

The study "Music Emotion Recognition with Standard and Melodic Audio Features" [Panda et al., 2015] introduces a fusion of standard and melodic audio features extracted from music recordings, leveraging Naïve Bayes, KNN and SVM algorithms for emotion classification. The dataset, built based on the All-Music knowledge base and mirroring the Music Information Retrieval eXchange (MIREX) Mood Classification task testbed, demonstrates the proposed approach's superiority over prior models relying solely on standard audio features.

Experimental results showcase remarkable performance, with the best outcome of a 64% F-measure achieved using SVM with just 11 features. The authors posit that incorporating standard and melodic audio features directly extracted from audio holds promise for further improving results. However, despite achieving high accuracy, the limited test collection included only three songs in the fear category.

Within the continuum of the 2015 paper on "Music Emotion Recognition with Standard and Melodic Audio Features," the research trajectory has shifted to confront the insufficiency of relevant musical characteristics for effective emotion identification [Panda et al., 2020b]. In response, the study introduces 29 innovative features tailored to refine emotion classification. To scrutinize the significance of these features, a new dataset, the 4QAED dataset (as discussed in Section 2.2), was crafted. The model proposed in a preceding work [Panda et al., 2015], underwent training with diverse features, culminating in peak performance. Combining the novel features with 71 standard features yielded an outstanding 76.4% F1-score. This achievement represents a noteworthy 9% improvement compared to the use of a baseline set of 70 features.

A distinctive method compared to those examined previously is the approach done by [Yang, 2021a], which proposes a model using NN technology that can analyze the entire music and accurately express the ups and downs of music emotion. The proposed model uses a combination of the BP neural network and Artificial Bee Colony (ABC) algorithm to extract features from the music and classify emotions. While the BP neural network is adequate for pattern recognition, it often struggles with getting stuck in suboptimal solutions, leading to less accurate results. The ABC algorithm, inspired by honeybee foraging behaviour, optimizes the initial weights and thresholds of the BP network, improving its ability to explore the solution space effectively. This combination enhances accuracy by avoiding suboptimal results and accelerates convergence, leading to faster and more reliable emotion detection across music tracks.

The dataset used in this research is the MediaEval Emotion in Music (MEM), splitting 80% of the entire data set to be used for training, and the remaining 20% is used for testing. The proposed model achieved an Root Mean Squared Error (RMSE) of 0.1322 for arousal and 0.1066 for valence.

Despite these results, the model disregards crucial high-level aspects such as lyrics and cultural context that significantly shape emotional responses to music.

3.2.2 Deep Learning Approaches

Deep learning has fundamentally changed how we understand and interpret the emotional content of music by autonomously extracting intricate features from raw data. Its proficiency in handling complex patterns and temporal dynamics, superior performance, and versatility position deep learning as an indispensable tool.

However, it's crucial to acknowledge that the efficacy of deep learning is contingent upon the specific task at hand and the quality and volume of the available data.

[Choi et al., 2016] pioneered the first DL approach for music auto-tagging, addressing emotion recognition. They introduced a Fully Convolutional Network (FCN) with four two-dimensional convolutional blocks, each comprising convolution, batch normalization, ReLU activation, and max pooling layers. This architecture processed spectrograms from raw audio signals, including STFT, MFCC, and Mel-spectrograms. The Mel-spectrogram, closely aligning with human auditory perception, yielded the best results. Their model produced a 50-feature binary vector for multi-label classification and achieved impressive Area Under the ROC Curve (AUC) scores of 0.894 on the GTZAN dataset and 0.851 on a subset of the MSD, setting a new benchmark in the field.

Following this work, [Yang, 2021b] presented three features, relative tempo, mean, and standard deviation of the average silence ratio, to model music emotion. Employing a NN classifier, they adeptly map the feature space to the emotion space, effectively categorizing emotions into happiness, sadness, anger, and fear.

Comprehensive experimentation on a corpus of 353 popular music pieces showcases remarkable results, boasting a precision of 67% and a recall of 66% in music emotion detection. The critical role of the BP neural network classifier is evident in its ability to map intricate features to nuanced emotions, thus laying the groundwork for subsequent advancements in the domain of DP for MER. While the initial results may seem promising, it is crucial to note that they can be misleading due to the considerable imbalance in the dataset used for evaluation.

Exciting findings were presented in the paper by [Gómez Cañón et al., 2021], which explores the correlation between speech and music in emotion recognition. In the study, the models were pre-trained using English and Mandarin speech, then fine-tuned with music excerpts labelled with emotion categories. The researchers found that features learned from the speech were transferable to

music for emotion recognition. Furthermore, the study demonstrated that using an intra-linguistic setting improved performance.

The researchers utilized the previously mentioned 4QAED dataset [Panda et al., 2018] to pre-train the models and employed a Sparse Convolutional Autoencoder (SCAE) to extract features from the speech data. They performed Bayesian optimization to select optimal learning rates and decays for the Adam algorithm and evaluated the models' performance using accuracy and F1 score metrics. Although the results may not have been satisfactory, the study confirmed a correlation between the language of speech and emotion in music.

MER has advanced significantly in recent years, with researchers exploring new and innovative ways to replicate humans' perception of emotions in music using ML techniques. One such approach is using RNN with LSTM units, which has shown promising results in predicting the continuous values of emotions on the axes of Russell's Circumplex model.

[Grekow, 2021] presents a novel approach to automatic emotion detection in music, using trained regression models to recognize emotions in music. It demonstrates the usefulness of dividing the data into sequences and using recurrent networks to achieve superior results compared to the SVM algorithm for regression. The study also analyzes the effect of the network structure and the set of used features on the results of the regressors recognizing values on two axes of the emotion model: arousal and valence. What sets this paper apart from others is its use of a segment length (6 seconds) different from the standard static MER and its proposal of a method of preparing data for recurrent neural networks by extracting various audio features from the music fragments. These features are then used to create sequential data for learning networks with LSTM units.

The research uses a database consisting of 324 six-second fragments of different genres of music: classical, jazz, blues, country, disco, hip-hop, metal, pop, reggae, and rock, taken from the publicly available GTZAN data collection. The best results were obtained with RNN, comprising two layers of 248 LSTM units. For arousal, the Mean Absolute Error (MAE) value was 0.12. For valence, the MAE value was 0.15. The results show that using a two LSTM layer RNN gives better results for both arousal and valence. One of the study's main limitations is that it can be challenging to determine which input features were used during the feature extraction process.

Paper	Approach	Emotion Taxonomy	Datasets	Features and Input	Models	Results	Notes/Observations
[Feng et al., 2003]	Classical ML	Four emotions: Happiness, sadness, anger and fear	223 popular songs	5 features	Feed-forward network	Accuracy values of 75% for sadness, 83% anger, 86% happiness and 25% for fear	Unbalanced dataset - fear class is poorly represented.
[Meyers, 2007]	Classical ML	Russell's A/V Model	372 songs	5 features (Mode, Harmony, Tempo, Rhythm, Loudness)	Decision Tree for classification and KNN for emotion prediction	Good results	No statistical method to assess the performance of the models.
[Panda et al., 2015]	Classical ML	MIREX's task 5 clusters	Based on MIREX taxonomy	11 features (9 melodic and 2 standard)	Naïve Bayes, KNN and SVM (top result)	0.64 F-measure	Limited test collection included only three songs in the fear category.
[Panda et al., 2020b]	Classical ML	Russell's A/V Model	4QAED	100 features (29 novel and 71 standard)	SVM	0.76 F-measure	The small, semi-automatically labeled dataset may have inaccuracies, and the model's performance is uncertain due to limited testing and lack of broader validation.

Paper	Approach	Emotion Taxonomy	Datasets	Features and Input	Models	Results	Notes/Observations
[Yang, 2021a]	Classical ML	Hevner emotional model	MEM	Mel-spectrogram	BP NN	RMSE of 0.1322 for arousal and 0.1066 for valence	The model overlooks key high-level factors such as lyrics and cultural context.

Table 3.2: Review of MER Classical ML Approaches.

Paper	Approach	Emotion Taxonomy	Datasets	Features and Input	Models	Results	Notes/Observations
[Choi et al., 2016]	Deep Learning	50 emotional tags	MagnaTag ATune and MSD	Mel-spectrogram, STFT and MFCC	FCN and variations of this architecture	0.894 AUC for MagnaTagATune and 0.851 AUC for the MSD	The use of the MSD undermines the credibility of the results, as the problem is treated as a multi-class classification task without adopting widely recognized emotion taxonomies.
[Yang, 2021b]	Deep Learning	4 labels (happiness, sadness, anger and fear)	223 samples retrieved from available sources	Relative tempo, mean, and standard deviation of the average silence ratio	NN	67% precision and 66% recall	Imbalance in the dataset used for evaluation
[Gómez Cañón et al., 2021]	Deep Learning	Russell's A/V Model	4QAED	Mel-spectrogram	CNN	0.48 F1-score	The study's results demonstrate a correlation between the language used in speech and the expression of emotion in music.

Paper	Approach	Emotion Taxonomy	Datasets	Features and Input	Models	Results	Notes/Observations
[Grekow, 2021]	Deep Learning	Russell's A/V Model	324 six-second fragments from the corresponding samples of GTZAN	529 features	Modified RNN and CNN	MAE of 0.12 for arousal and 0.11 for valence	Difficulty in identifying which input features were utilized during the feature extraction process.

Table 3.3: Review of MER Deep Learning Approaches.

Tables 3.3 and 3.4 present studies on MER using Classical ML and DL approaches. Different emotion models are used, ranging from basic emotions to Russell's Circumplex model.

Various datasets are used in these studies, including popular songs and specific collections like DEAM and 4QAED. It's worth noting that the choice of dataset can significantly impact the results obtained. For instance, two studies conducted [Panda et al., 2015, 2020b] showed a 12% improvement in the F1-score metric due to using the 4QAED dataset, which was more balanced across the different Russell quadrants than the other dataset.

The models exhibit varying performance, with certain ones achieving high accuracy or F1-scores while others have lower performance. It is important to note that different studies may use different metrics to evaluate their models, making direct comparisons challenging.

However, many studies note limitations or challenges in their work, such as unbalanced datasets, overlooking key factors like lyrics and cultural context, and difficulties in feature extraction. These observations highlight the complexity of MER.

3.3 Music Emotion Variation Detection

As previously mentioned, unlike static MER, the focus of Music Emotion Variation Detection (MEVD) is to analyse emotion variation throughout songs. This section delves into the realm of MEVD approaches, presenting a comprehensive review of the latest advancements in the field. It identifies the challenges associated with these approaches and provides valuable insights for implementing a robust architecture and methodology.

3.3.1 Classical Approaches

[Schubert, 2004] introduced a method that used linear regression models to predict the emotional content of a song in terms of arousal and valence. This model relied on five standard audio features: melodic contour, tempo, loudness, texture, and spectral centroid. Annotations were collected from 67 participants based on Russell's A/V model to create the dataset required to train and evaluate the model. The dataset consisted of four Romantic music pieces and annotations made in one-second intervals.

The evaluation of the model revealed that variations in loudness and tempo correlated with changes in arousal, but none of the analyzed features significantly affected valence. It is important to note that this research uses a limited dataset of songs from one genre.

Another approach done by [Panda and Paiva, 2011] proposes a system that predicts the emotion of small segments from full songs based on various audio

features, using SVMs for classification and regression analysis and the models trained with music clips previously annotated with A/V values.

The study used 189 clips, each lasting 25 seconds, from various genres, mainly Pop/Rock from Western and Asian artists. They segmented the clips using two audio frameworks, Marsyas and MIR Toolbox.

The researchers recruited volunteers to annotate the changes between quadrants in 57 complete songs to conduct the testing. Two volunteers annotated each song. However, only the songs with an 80% matching rate between both annotators were selected for testing, resulting in the shortening of the testing set from 57 to 29 songs.

The study found that the accuracy of the system varies depending on the quadrant of Thayer's model of emotion, with higher accuracy for segments in the first and fourth quadrants, which correspond to positive valence, and lower accuracy for segments in the second and third quadrants, which correspond to negative valence.

The results obtained an average of 53.71% in terms of accuracy. Volunteers' annotations may be inconsistent and subjective, affecting the system's accuracy.

[Markov and Matsui, 2015] explored Gaussian Processes (GP) in recognizing human emotions in speech and music, showcasing their superiority over other models in capturing nonlinear data relationships. The study utilized the "MediaEval'2014" database consisting of excerpts from 1744 songs belonging to various genres. The researchers randomly selected 500 clips for training and another 500 for testing. The results showed an RMSE of 0.0972 for arousal and 0.1002 for valence. Despite the successful results, the authors acknowledge the computational complexity of using GPs in real-time applications.

3.3.2 Deep Learning Approaches

A recent study by [Malik et al., 2017] builds upon the work of [Choi et al., 2016] by employing a Convolution Recurrent Neural Network (CRNN) architecture for music emotion recognition. This approach combines convolutional layers for feature extraction with recurrent layers for capturing temporal dependencies, enabling effective prediction of emotions within the two-dimensional A/V space defined by the Russell model.

The study used 431 audio samples with a duration of 45 seconds for training the model. However, only the final 30 seconds of each sample were used for training. The annotations were made every 500 ms with arousal and valence values in the range of [-1,1], resulting in 60 annotations for each of the 30-second samples. The evaluation used 58 songs from the MedleyDB dataset [Bittner et al., 2014] and the music website Jamendo.

Interestingly, the system using baseline features achieved the best results, with an RMSE of 0.202 for arousal and 0.268 for valence.

A Bidirectional Convolutional Recurrent Sparse Network (BCRSN) model was proposed by [Dong et al., 2019], consisting of a robust model that combines the strengths of CNNs and RNNs for music emotion recognition using Russell's model.

The BCRSN model is designed to learn different levels of features adaptively through convolution and subsampling operations. It leverages the feature maps obtained by CNN as the input of RNN to enhance the model's prediction performance, incorporating a Weighted Hybrid Binary Representation (WHBR) method to reduce computational complexity.

A portion of the DEAM dataset, consisting of 431 complete songs, was utilised to train this model. The assessment set contains 58 songs from the same database. The researchers also used 240 pop songs from the MTurk dataset 7 to assess the model's generalisation ability. The results on the DEAM dataset show a RMSE of 0.123 for valence and 0.102 for arousal. On the MTurk dataset, the RMSE for valence is 0.145, and for arousal, it is 0.079.

The proposed architecture by [Orješek et al., 2022] consists of stacking a one-dimensional CNN layer, a distributed layer with autoencoder-based iterative reconstruction for latent feature extraction, followed by bidirectional Gated Recurrent Unit (GRU), and the max out fully connected layer for efficient valence-arousal regression from the latent features that mines emotion-related features from the raw audio waveform.

The datasets used for training and evaluation were used in [Dong et al., 2019], and the emotional taxonomy was also Russell's Circumplex model.

The experimental results show an RMSE for valence of 0.114 and arousal of 0.105 that outperforms the BCRSN approach trained with Mel-spectrogram as input, which performs better in arousal, and the LSTM-RNN architecture outperforms the proposed system in valence.

Paper	Approach	Emotion Taxonomy	Datasets	Features and Input	Models	Results	Notes/Observations
[Schubert, 2004]	Classical ML	Russell's A/V Model	4 romantic songs, annotated every 1 second	5 features (melodic contour, tempo, loudness, texture, and spectral centroid)	Linear regression models	Detected changes in arousal but not in valence	Small dataset and comprised by only one genre.
[Panda and Paiva, 2011]	Classical ML	Russell's A/V Model	57 full songs annotated in 25 second intervals	Standard audio features, such as timbre and rhythm	SVM	53.71% accuracy	Small dataset.
[Markov and Matsui, 2015]	Classical ML	Russell's A/V Model	MediaEval 2014, annotated in 0.4 second intervals	Standard audio features	Gaussian Process regression	RMSE of 0.0972 for arousal and 0.1002 for valence	Computational complexity of this model makes it difficult to apply in large-scale applications

Table 3.4: Review of MEVD Classical ML approaches.

Paper	Approach	Emotion Taxonomy	Datasets	Features and Input	Models	Results	Notes/Observations
[Malik et al., 2017]	Deep Learning	Russell's A/V Model	DEAM 431 sample subset for training and 58 complete songs for evaluation	Standard features or spectrogram	CNN and RNN	RMSE of 0.202 for arousal and 0.268 for valence	The length of the samples may evoke mixed emotions
[Dong et al., 2019]	Deep Learning	Russell's A/V Model	Portion of the DEAM dataset and MTurk	Spectogram	BCRSN	RMSE of 0.101 for arousal and 0.123 for valence in the DEAM dataset; RMSE of 0.079 for arousal and 0.145 for valence in the MTurk dataset	Model is very complex leading to long training time
[Orješek et al., 2022]	Deep Learning	Russell's A/V Model	Portion of the DEAM dataset	Standard audio features and Mel-spectrogram	Modified RNN and CNN	RMSE for the valence of 0.114 and arousal of 0.105	Questioning the model's performance arises from comparing its RMSE scores to those of other models.

Table 3.5: Review of MEVD Deep Learning Approaches.

Tables 3.5 and 3.6 present studies on detecting variations in music emotions using Classical ML and deep learning. Both approaches have their strengths and weaknesses.

Classical ML models, such as Linear Regression, SVM, and Gaussian Process Regression, are relatively simple and computationally efficient. They are easier to interpret and can provide insights into the most important features of emotion recognition. These models may not capture complex, non-linear relationships in the data.

DL models, such as CNNs, RNNs, and hybrid models, can model complex, non-linear relationships and capture high-level features in the data. However, these models are computationally intensive and require large amounts of data. Additionally, these models can be prone to overfitting, especially when the dataset is small or unbalanced.

The diversity and size of datasets used are often limited, raising questions about the generalizability of the results. Using larger and more diverse datasets could improve the robustness of the findings.

To sum up, Classical ML and DL have advantages and disadvantages. The selection between the two should depend on the specific requirements of the task, such as the complexity of the data, the availability of computational resources, and the need for model interpretability. In the future, researchers should focus on developing models combining both approaches' strengths, such as interpretable deep learning models. Additionally, they should work on devising more reliable and comprehensive evaluation techniques.

3.4 Segmentation Tools

Segmentation tools are essential in MER and MEVD fields. These tools help to break down audio tracks into distinct segments, each representing a meaningful unit within the musical composition. This process involves identifying boundaries, transitions, and structural elements, contributing to a detailed understanding of the music's emotional dynamics. Segmentation tools are helpful for researchers, musicians, and technology developers in various applications, from analyzing emotional signals in music to enhancing the accuracy of emotion-related models. This exploration will explore the significance of segmentation tools, their structures, and their impact on advancing our comprehension of musical emotions.

3.4.1 DeepChorus

The DeepChorus model comprises two key components: a Multi-Scale Network for generating initial representations of chorus segments and a Self-Attention Convolution Network for further processing the features into probability curves that indicate the presence of chorus. The model then employs an adaptive thresh-

old to convert the output probability curve into a binary value, i.e., chorus or non-chorus. For a better understanding, refer to Figures 3.3 and 3.4, which visually represent the model.

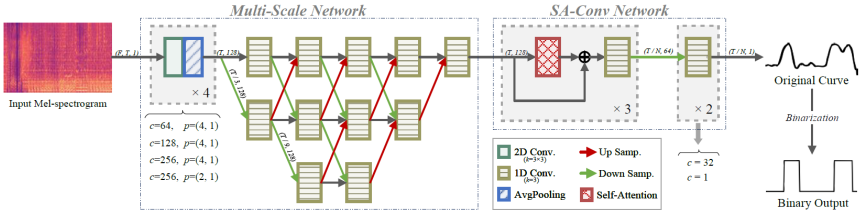


Figure 3.3: Visualization of the DeepChorus Model [He et al., 2022].

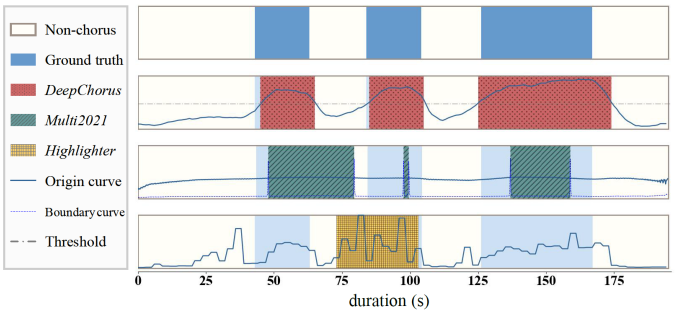


Figure 3.4: Classification of Chorus using an adaptive threshold.

The study employed a training dataset of 886 tracks from the HARMONIX dataset [Nieto et al., 2019] and 102 songs from The Beatles and Michael Jackson sourced from the Isophonics dataset [Mauch et al., 2009]. Diverse datasets were tested, including 508 songs from the SALAMI [Smith et al., 2011], 100 pieces from RWC datasets [Goto et al., 2002], and 210 songs with the "Popular" tag from SALAMI, specifically chosen to focus on widely recognized tracks that are more likely to exhibit standard structural features. This comprehensive approach allowed for a thorough evaluation of the proposed method's effectiveness and generalizability. This model achieved F1-scores of 0.501, 0.675, and 0.611 on the three test datasets above.

3.4.2 All-in-One

The development and success of the All-in-One segmentation tool [Kim and Nam, 2023] have certainly been promising in music analysis. By comprehensively analysing all structural elements within a song, All-in-One has set itself apart from its counterparts and shown great potential for further advancements in the field. Prior efforts, including works by [Ullrich et al., 2014] and [Wang et al., 2022], laid the groundwork for song segmentation. However, the results achieved by All-in-One have surpassed those of previous methods.

Figure 3.5 represents the All-in-One structure where the model first extracts features and effectively reduces dimensionality from demixed sources. The demix-

ing process uses the Hybrid Transformer Demucs source separation algorithm, which extracts drum, vocal, bass, and other instrument sources. The extracted sources are then fed into three convolutional and max pooling layers, resulting in a compact representation of the separated stems.

To model both temporal and instrumental dependencies, the model stacks 11 Transformer modules. Each module comprises a 1D Dilated Neighborhood Attention (DiNA) block and a 2D Neighborhood Attention (NA) block. The 1D DiNA block captures long-term temporal dependencies, and the 2D NA block focuses on inter-instrument dependencies.

The model employs four fully connected layers to predict probabilities for beat, downbeat, segment boundary, and structure label for each time frame. Post-processing steps such as applying Dynamic Bayesian Networks (DBN) for beat and downbeat tracking, refining probabilities, and selecting peaks enhance the model's accuracy. A peak-picking method based on sliding window averages and highest probabilities are employed for segmentation and structure labeling.

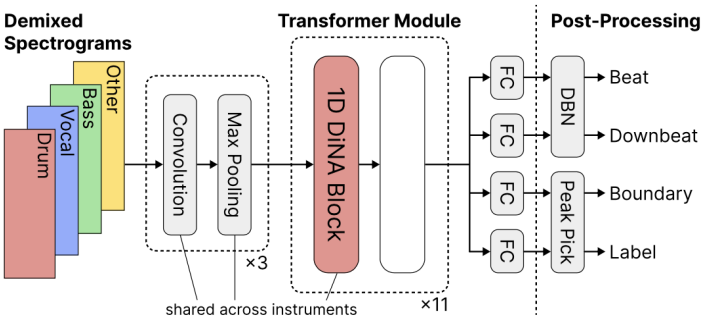


Figure 3.5: Visualization of the All-in-One Model [Kim and Nam, 2023].

The Harmonix Set [Nieto et al., 2019], consisting of 912 popular Western songs with annotations for beats, downbeats, and functional segments, was used to train and evaluate the model. The paper has adopted the data preprocessing steps recommended in the previous work of [Wang et al., 2022] to guarantee data consistency and compatibility. For performance evaluation, the authors employ an 8-fold cross-validation strategy. Within these folds, six are allocated for training, one for validation, and one for testing.

In the context of segmentation, the proposed model exhibited notable performance, achieving a hit rate F-measure of 0.660 with a 0.5-second time window. Moreover, the model's performance in structural labeling is remarkable as it has achieved state-of-the-art results. Compared with the ground truth, the predicted structural segmentation of pairs of frames has an F-measure of 0.738, demonstrating its high accuracy.

3.5 Summary

Chapter 3 thoroughly explores the current state of the art in MER and MEVD models. It begins with an in-depth discussion of MER datasets, emphasizing their crucial role in advancing MER and MEVD research. Static MER focuses on identifying static emotions in music by analyzing smaller excerpts, whereas MEVD aims to understand emotional variations over time in entire songs, necessitating comprehensive song annotations.

The chapter highlights several challenges faced by current datasets, such as limited size, diversity, and annotation inconsistencies. These limitations impact the reliability and applicability of MER models. The subjective nature of emotions, the inconsistency in annotations, and imbalances in class distribution are significant hurdles. Additionally, capturing temporal dynamics and ensuring privacy in datasets present ongoing complications, copyright restrictions, lack of standardization, and domain specificity. To address these issues, researchers must develop more comprehensive and high-quality datasets.

Following this, chapter 3 examines static MER and MEVD approaches. Classical methods involve feature extraction from music, focusing on tempo, pitch, timbre, and rhythm. These features are then used to train ML algorithms to classify music's emotional content. For instance, in the study [Feng et al., 2003], researchers used neural networks to classify emotions in music, achieving notable accuracy for emotions like happiness, sadness, and anger.

Furthermore, this chapter also delves into deep learning approaches, which have significantly transformed MER and MEVD by autonomously extracting intricate features from raw data. Deep learning's proficiency in handling complex patterns and temporal dynamics makes it an indispensable tool for MER and MEVD. Notable studies are discussed, such as the work of Panda and Paiva in 2011, which employed an SVM model for automatic emotion tracking in music. This study utilized Russell's Circumplex model to predict the emotional content of music over time, highlighting the potential of ML algorithms in capturing the dynamic nature of music emotions.

Finally, the chapter then shifts focus to the significance of segmentation tools in MER and MEVD, explaining how tools like DeepChorus and All-in-One help break down audio tracks into distinct segments, each representing a meaningful unit within the musical composition, thus aiding in detailed analysis. This segmentation enables a more detailed analysis of the structural and emotional elements of music.

Chapter 4 builds upon understanding the approaches and methodologies used in MER and MEVD by detailing the experimental procedures and outcomes. This upcoming chapter will focus on implementing the All-in-One segmentation tool and applying methodologies like SVMs and CNNs to enhance the classification accuracy of emotional segments in music.

Chapter 4

Methods and Experiments

In Chapter 4, the research methodologies discussed earlier are applied. The process begins by replicating previous studies to establish a reliable baseline. Then, experiments using segmentation tools, including exploration of the DeepChorus tool, are conducted. However, only the All-in-One tool is used to segment the songs for further use in Classic and DL approaches. Classical and deep learning approaches are tested, and a detailed analysis of the implementation, datasets, and results is provided.

This chapter details the experiments conducted to evaluate whether the All-in-One tool delivers effective results in ML and deep learning. The goal is to test its effectiveness and understand how well it helps for more accurate emotion variation detection.

4.1 Replication of Previous Work

This section describes the work done by Panda and Paiva to predict A/V values and quadrants, classifying emotion in window segments from full songs. Next, an approach to solve this problem is presented, trying to replicate the study mentioned above.

4.1.1 Previous Work

In 2011, Panda and Paiva studied MEVD using SVMs and audio features. To this end, the authors used a dataset of 189 clips originally collected by [Yang et al., 2008] for training, mainly Pop/Rock from Western and Asian artists, each 25 seconds long, annotated with arousal and valence values by volunteers.

The authors followed two different approaches to achieve categorical MEVD, i.e., predict quadrants: one based on classification, where they trained a Support Vector Classification (SVC) to predict the quadrant of each audio segment, and another based on regression, where two Support Vector Regression (SVR) were

trained to predict the A/V values of each segment, which were later converted to quadrants of the Russell's Circumplex.

From the 189-clip dataset, various audio features were extracted from each clip using two audio frameworks, Marsyas and MIR Toolbox. This was followed by Forward Feature Selection (FFS) to select the most relevant features for this specific approach. The researchers experimented with various combinations of features, window sizes, and frameworks to obtain the classifier or regressors that better predicted the quadrant of each clip.

Next, the authors used the two emotion tracking solutions to track emotion variation over time by predicting the quadrant of 29 complete song versions segmented into small consecutive 1.5-second clips. The 29 songs were selected from an initial set of 57 analyzed by volunteers, only including those achieving an 80% or higher matching annotation rate. They compared the predicted emotion tracks with manual annotations made by two volunteers, as seen in Figure 4.1, and measured the accuracy of their approach. Table 4.1 presents the results of this study.

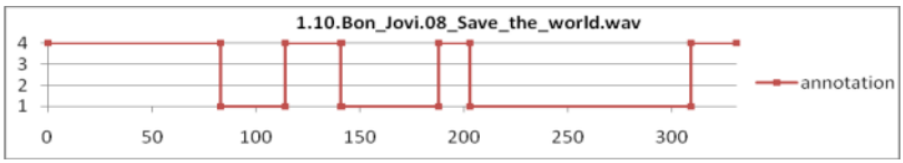


Figure 4.1: Example of a tracking annotation [Panda and Paiva, 2011].

	All features		Feature selection	
	Quadrants	A/V	Quadrants	A/V
Marsyas	53.45%	48.90%	52.55%	50.89%
MIR Toolbox	52.70%	55.95%	54.51%	56.30%
Marsyas + MIR T.	53.66%	55.95%	54.72%	54.96%

Table 4.1: Comparison of performance between Marsyas, MIR Toolbox, and their combination using all features and feature selection in the 29 song dataset.

4.1.2 Replication of Previous Work

In order to explore and validate the MEVD methodologies in emotion tracking, this subsection begins by replicating the work of [Panda and Paiva, 2011] on MEVD using SVMs and audio features. A vital aspect of this endeavour entailed the conversion of the source code from MATLAB to Python. This conversion was carried out to enhance the accessibility and user-friendliness of the methodology for researchers interested in exploring diverse MEVD techniques on audio using different models.

Contrasting with the original study, two feature files were tested independently, one with the features extracted using the Marsyas tool and another using the

MIR Toolbox. Thus, the tests were divided into three parts: prediction of emotion using the MIR Toolbox, prediction of emotion using the Marsyas tool, and prediction using both Marsyas and MIR Toolbox.

Similarly to the approach used in [Panda and Paiva, 2011], an SVC was trained to predict the quadrant of each segment of an entire song. Additionally, a regression-based approach was employed, where two SVR were trained to predict the A/V values of each segment. These A/V values were then converted into quadrants of Russell's Circumplex model and used as labels for classification.

In pursuit of improved classification results, the SVM model needed further optimization regarding the quadrants. As for A/V, the results were satisfactory, and there was no need to further optimize the model since the aim of this work was to replicate the results of [Panda and Paiva, 2011]. The distribution of songs in Yang's dataset, used for training, across the quadrants was imbalanced, with an excess in the first quadrant and a shortage in the second quadrant, as seen in Figure 4.2.

To address this issue, we experimented by adjusting the SVM model to handle unbalanced data more effectively. This involved incorporating class weights that penalized the misclassification of the minority class more heavily. However, while managing the imbalanced song distribution, these adjustments to the SVM inadvertently resulted in overfitting. Consequently, this overfitting limits the model's ability to generalize well to new, unseen data, exhibiting poor performance and failing to achieve satisfactory results.

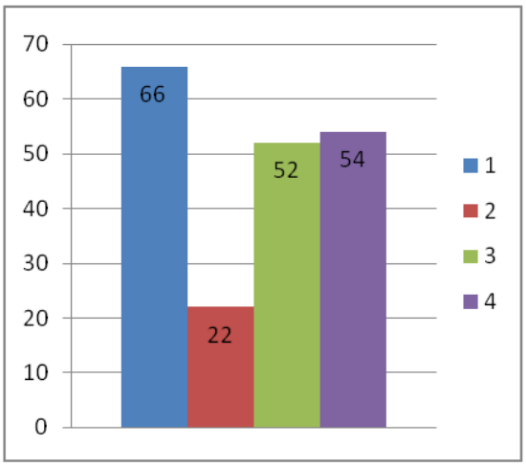


Figure 4.2: Yang's dataset distribution across quadrants [Panda and Paiva, 2011].

Various optimization techniques, including grid search, bayesian search, and random search, were employed to identify the optimal hyperparameters that would work best with our unique data distribution.

The hyperparameters that demonstrated peak performance for SVM quadrants, both with and without FFS, across the different feature types are as follows (Table 4.2):

	FFS			No FFS		
Approach	**C**	**gamma**	**kernel**	**C**	**gamma**	**kernel**
MIR	49.835	0.017	`poly`	13.679	5.674	`RBF`
Marsyas	1.090	0.012	`RBF`	1.047	7.702	`linear`
MIR+Marsyas	0.214	0.944	`RBF`	1.020	0.013	`RBF`

Table 4.2: Selection of the hyperparameters for the SVM model.

Since MEVD Panda dataset was too imbalanced, with 10 songs in the first quadrant, 7 in the second, only 2 in the third, and 10 in the fourth quadrant, the MEVD dataset was created for testing. This dataset was based on the original 29 songs but added five more songs to the third quadrant to reduce the imbalance. The MEVD dataset contained 10 songs in quadrant 1, 7 in quadrant 2, 7 in quadrant 3, and 10 in quadrant 4.

Finally, the predicted emotions were compared to the ground truth annotations, the results are presented in Table 4.3.

	All features		Feature selection	
	Quadrants	A/V	Quadrants	A/V
Marsyas	50.36%	55.04%	47.63%	55.03%
MIR Toolbox	51.58%	55.11%	49.48%	55.10%
Marsyas + MIR T.	55.15%	55.23%	55.61%	55.53%

Table 4.3: MEVD tracking results for each framework and the combination of the framework in the MEVD datatset.

4.2 Experiments with Segmentation Tools

In this section, we will delve into segmentation tools. Segmentation tools, such as DeepChorus [He et al., 2022] and All-in-One [Kim and Nam, 2023], are software or algorithms designed to partition a musical piece into distinct sections based on criteria such as musical structure. The primary objective of these tools is to enable a more focused analysis of specific portions of the song, allowing researchers to examine and comprehend the characteristics, patterns, or emotions associated with each segment more effectively.

In this case, the hypothesis was to test whether these segmentation tools, which divide the song based on a more logical musical structure rather than a fixed division of 1.5 seconds, could improve emotion prediction results within these segments.

4.2.1 DeepChorus

DeepChorus is a deep learning model created to detect the chorus in music. It identifies the most repeated and recognizable song sections, usually the chorus,

by analyzing Mel-spectrograms using a combination of multi-scale convolution and self-attention mechanisms.

Preliminary experiments conducted with the DeepChorus tool tested its performance across several music genres, including Pop, Rock, Hip-Hop, Latin, Electronic, and Country, with the objective of identifying and understanding the tool's weaknesses. With these, we concluded that DeepChorus struggles to provide accurate predictions, mainly when a pre-chorus is present in a song.

The code was modified to visualize the comparison between the ground truth chorus (displayed in the top graph of the image) and the predicted chorus (shown in the bottom graph).

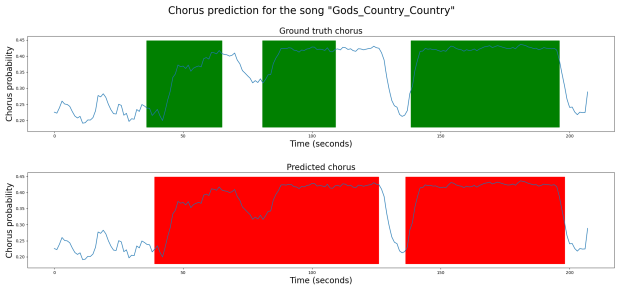


Figure 4.3: Chorus identification for country music "God's Country".

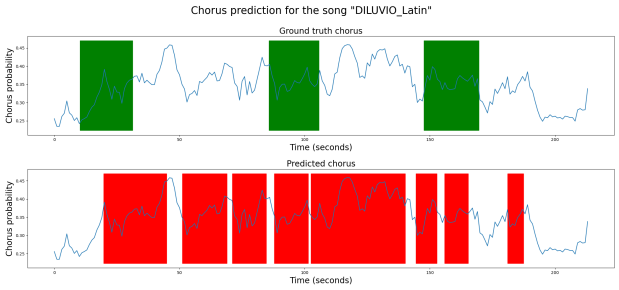


Figure 4.4: Chorus identification for latin music "Diluvio".

DeepChorus is less effective for music genres like Latin, as seen in Figure 4.4 and Electronic, but yields satisfactory results for Country, as seen in Figure 4.3, and Hip-Hop. This discrepancy can be attributed to Country and Hip-Hop music typically following a more standardized structure, featuring an intro followed by a chorus, verse, and a recurring pattern. On the contrary, Latin music deviates from this structure, incorporating elements like pre-chorus, bridge, and other structural components that DeepChorus may struggle to analyze effectively.

Due to its limitation of only detecting the chorus, DeepChorus was not used in further experiments with the classical and deep learning approaches. This limitation prevents the tool from taking advantage of the structural separation within a song to detect emotional variation, as it focuses solely on identifying the chorus and cannot segment other important sections like verses, bridges, and interludes. This restriction reduces its potential for more detailed emotional analysis across

the entire musical structure, making it unsuitable for the more comprehensive segmentation needed in these experiments.

4.2.2 All-in-One

The All-in-One model is a deep learning-based tool for comprehensive music structure analysis. It performs tasks such as beat and downbeat tracking, segmentation of a song into distinct sections, and labeling of segments based on their roles, such as identifying verses and choruses. Using advanced transformer architectures with dilated neighbourhood attention mechanisms, the model effectively captures both local details and long-term patterns in the musical piece. This enables segmentation by detecting where one section ends and another begins and labeling by understanding the context of each segment within the overall song structure.

The experiments conducted in this section aimed to better understand the All-in-One tool's capabilities across the same music genres as DeepChorus and gain insights into its potential limitations. We found that this tool faces challenges in delivering accurate predictions, particularly in songs that include a pre-chorus, post-chorus, interlude, and bridge.

According to the analysis, All-in-One, like DeepChorus, has demonstrated superior performance to Country songs that follow the traditional musical structure of chorus, verse, chorus, verse. However, All-in-One's performance was comparatively weaker in Latin music, with structural elements such as pre-chorus, post-chorus, interlude, and bridge. A script was created to visualize the comparison between the ground truth segments (displayed in the top graph of the image) and the predicted chorus (shown in the bottom graph).

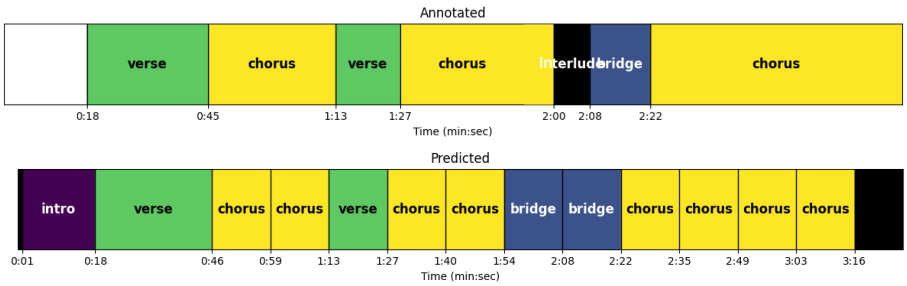


Figure 4.5: Segment and segment label identification for country music "God's Country".

From the two above Figures, it is conclusive that All-in-One also faces difficulties similar to DeepChorus, as it has poor results in Latin music as seen in Figure 4.6 and better results in country music as seen in Figure 4.5. Like DeepChorus, the results are better in Hip-Hop and Country music and much worse in Latin and Electronic music. Music genres that got worse results may be because of structural elements like pre-chorus, post-chorus and bridge.

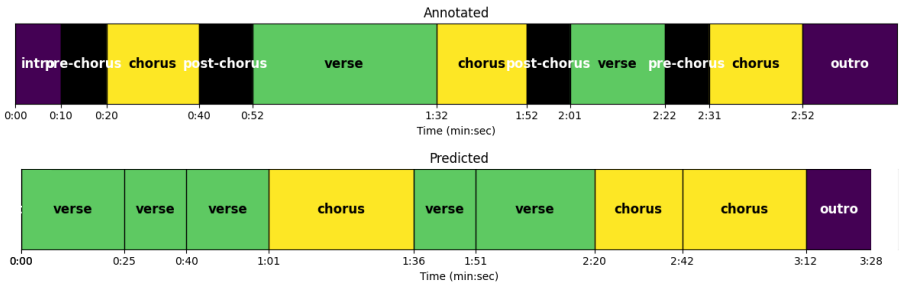


Figure 4.6: Segment and segment label identification for latin music "Diluvio".

4.3 Classical Approach

This section explores alternative segmentation strategies in MEVD, specifically comparing variable segments based on musical structure obtained via All-in-One with fixed-size 1.5-second segments, as previously used in [Panda and Paiva, 2011]. The primary objective was to assess whether a segmentation approach grounded in musical structure could provide advantages over a 1.5-second fixed-interval method. In addition to the 2011 study, this research also tested an additional set of emotionally relevant features proposed by Panda et al., which include musical dimensions such as melody, harmony, rhythm, dynamics, expressivity, and texture, in comparison to the standard audio features that are primarily based on low-level spectral characteristics [Panda et al., 2020a]. This approach allowed for a more comprehensive analysis of how different segmentation and feature sets influence the accuracy of emotion classification in music.

The aim is to identify the optimal segmentation strategy and segment size for classifying emotions into Russell's four quadrants using the MERGE Audio Complete dataset, proposed in [Louro et al., 2024b], for optimisation purposes. The dataset includes 3,554 clips, each 30 seconds long. The distribution of the dataset is depicted in Figure 3.1. The test set used for our experiments was the MEVD dataset, which, as previously mentioned in Section 4.1.2, is based on Panda's 29-song dataset, with the addition of 5 songs specifically in the third quadrant to help reduce the imbalance of the dataset. This resulted in 34 full-length songs, as shown in Figure 4.7. The emotion annotation for the training clips is static, with no variation over time. Additionally, a 3-fold cross-validation experiment with 30 repetitions (30x3-fold CV) on the test dataset was conducted to evaluate the model's performance on a dynamic dataset where emotions vary over time.

The achieved results, with the best F1-score of 53.17%, indicate that the potential of this song segmentation approach was not fully utilized, possibly due to the small dataset used. Nonetheless, the preliminary results are encouraging.

4.3.1 Datasets

The classical approach utilized the MERGE Audio Complete dataset for training, as described in Section 3.1, and the MEVD dataset for evaluation, as described in

Section 4.1.2. An illustration of the MEVD dataset can be seen in Figure 4.7.

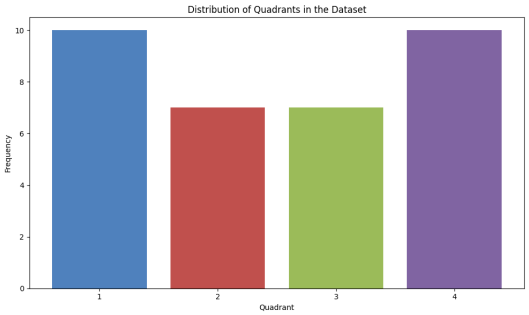


Figure 4.7: MEVD dataset distribution across quadrants.

4.3.2 Methodology

This subsection covers the methodology used in the entire process, from feature extraction to classification. Figure 4.8 gives an overview of the complete process of classifying emotion using an SVM model. Beyond the depicted below, the features were previously pre-processed and selected for the experiment to train the model and optimise the parameters.

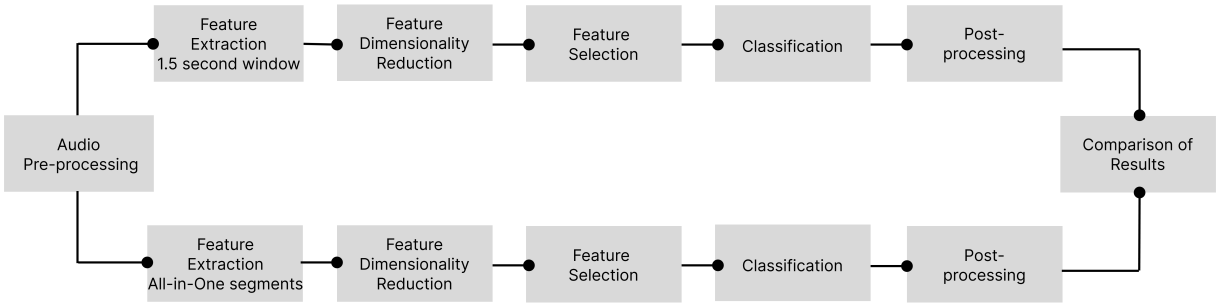


Figure 4.8: High level overview of the methodology.

Feature Extraction with 1.5-second segments

The feature extraction process involved standardising all audio clips from the dataset to a predefined format - specifically, the WAV PCM format with a sampling rate of 22050 Hz, 16-bit quantisation, and mono-aural configuration.

According to studies conducted by Vieillard et al., it takes an average of 483 ms, 1446 ms, 1737 ms, and 1261 ms to recognize happy, sad, scary, and peaceful excerpts, respectively. Emotions such as scary and peaceful, associated with the third and fourth Russell's quadrants, tend to take more time to detect than emotions like happiness. This is because emotions from the third and fourth Russell's quadrants are more complex and have less energy, making them more difficult to detect.

However, a more recent study by Paquette et al. found that listeners accurately identified emotions associated with brief musical clips averaging 1.5 seconds.

Therefore, for the MEVD dataset, features were extracted from 1.5-second windows within each song after standardization. In contrast, features were extracted from each 30-second clip for the MERGE Audio Complete dataset, as it was not feasible, due to the computational cost, to divide the 3,554 MERGE Audio Complete samples into 1.5-second windows.

Feature Extraction with All-in-One segments

The feature extraction uses the segments predicted by the segmentation tool. The window size is dynamic and varies based on the segment size. An example of song segmentation is provided in Figure 4.9, where the window size varies based on each segment size, where the first segment has a length of 21 seconds and the second has 16 seconds.

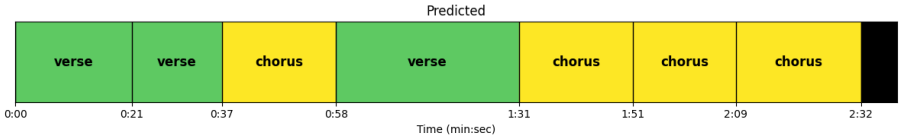


Figure 4.9: Segment and segment label prediction for the music "Tell Laura I Love Her".

It is important to evaluate the performance of the segmentation tool before using its predicted segments for feature extraction. The dataset used for this evaluation consists of 34 musical pieces covering all four emotional quadrants. To annotate these 34 songs regarding structural segments, we consulted the Genius website (https://genius.com/), which provided detailed insights into the song structures. This process ensured the segmentation aligned accurately with each song's composition and lyrics. An in-depth analysis of the metrics employed in [Kim and Nam, 2023] was undertaken to conduct this evaluation.

The authors of the All-in-One paper mentioned specific metrics used for segmentation evaluation. For segmentation assessment, they used the F-measure of hit rate at 0.5 seconds, and for evaluating segment labelling, they used the F-measure of pairwise frame-level clustering. However, [Nieto et al., 2020] is referenced for detailed insights into these metrics.

The hit rate metric assesses the accuracy of predicted segment boundaries against annotated boundaries by quantifying the proportion of correctly identified boundaries within a predefined tolerance parameter (typically 0.5 seconds). It combines precision and recall, representing the ratios of correctly identified boundaries to predicted and annotated boundaries, respectively, to compute the F1-measure, providing a comprehensive evaluation of segmentation accuracy. To assess the overall F1-measure of the dataset, a weighted average F1-measure was employed.

In contrast, the Pairwise Clustering metric for segment labelling compares pairs of time frames with identical labels within a given segmentation. If two frames

are labelled "A," they form a pair. This metric computes the precision and recall of the estimated pairs against the reference pairs and merges them using the F1-measure. A higher F1-score denotes better agreement between the real and predicted segmentations, thus indicating better segment labelling performance.

An evaluation was conducted on this dataset using these two particular metrics. The analysis resulted in a weighted average F-measure of 70.10% for the segmentation and a Pairwise Clustering score of 73.56% for the labelling of these segments, indicating excellent performance. Therefore, although not perfect, the predicted segments can be used for feature extraction.

Dimensionality Reduction

After feature extraction, a dimensionality reduction step was performed. Initially, features with zero standard deviation were removed, followed by eliminating heavily correlated features with a correlation factor exceeding a predetermined threshold (set experimentally at 0.9). When two features exhibited high correlation, the logic was to eliminate the second feature of the correlated pair.

Feature Selection

The RelieF algorithm is used to select the most essential features. The algorithm ranks features based on their importance, after which each set of top features is further analysed. Varying subsets of features were selected, denoted by the parameter X, ranging from 5 to 1250 integer values.

Training Phase

Hyperparameter optimization was carried out on each set of top features using a Bayesian search to customize optimal SVM models for specific datasets. This comprehensive approach encompassed parameters such as kernel type, gamma, cost, and polynomial degree (if applicable), covering a range of values for cost (1e-6 to 5000) and gamma (1e-6 to 100) and polynomial degree (1 to 5). The kernel types considered included Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid, as outlined in [Brownlee, 2024].

The optimization process utilized repeated stratified 10-fold cross-validation with ten repetitions, aligning with the approach proposed by [Duda et al., 2001]. This methodology aimed to maximize the F1-score for each dataset, thereby improving the robustness and effectiveness of the classical SVM-based music analysis framework.

A test was conducted on a range between 5 to 1250 features to determine the optimal number of features for optimal model performance. The conclusion was that the model performs best when using 900 features. Table 4.4 displays the F1-scores achieved by the model across different numbers of features and hyperparemeters.

Number of Features	Cost	Gamma	Kernel	F1-score
5	3303.03289	0.14883	RBF	0.524
10	4999.99999	0.02128	RBF	0.583
20	4999.99999	0.00734	RBF	0.603
30	4999.99999	0.00296	RBF	0.623
40	4999.99999	0.00204	RBF	0.644
50	4999.99999	0.00163	RBF	0.653
60	4087.42455	0.00079	RBF	0.653
70	4999.99999	0.00053	RBF	0.660
80	1162.23762	0.00086	RBF	0.657
90	71.81172	0.00268	RBF	0.656
100	83.36123	0.00164	RBF	0.664
150	17.99398	0.00225	RBF	0.677
200	24.16841	0.00093	RBF	0.685
250	18.94706	0.00059	RBF	0.686
300	9.63237	0.00086	RBF	0.694
400	53.74163	0.00025	RBF	0.699
500	8.29674	0.00088	RBF	0.703
600	4.83851	0.00085	RBF	0.705
700	3.65718	0.00064	RBF	0.708
800	3.67147	0.00057	RBF	0.710
900	**4.96317**	**0.00049**	**RBF**	**0.713**
1000	3.94264	0.00045	RBF	0.712
1100	5.53133	0.00030	RBF	0.711
1150	2.34943	0.00066	RBF	0.711
1250	4.29207	0.00029	RBF	0.712

Table 4.4: Global F1-scores for different numbers of features and different SVM Hyperparameters.

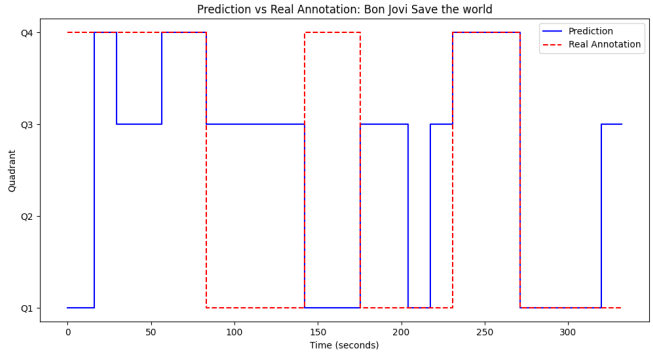


Figure 4.10: Comparison between ground truth annotations and model predictions for variable segments using standard features.

Segment Prediction

After training, the computed hyperparameters are applied to the model, which is then used to predict the class of each fixed or variable segment. Figure 4.10

compares the actual annotations with the model's predictions for the variable segments.

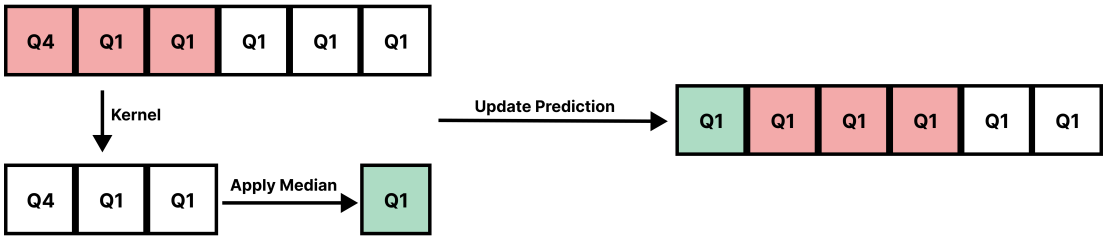


Figure 4.11: Concept of median filtering in data processing.

Post-Processing

After conducting tests, it is crucial to perform post-processing to identify and remove potential outliers. This step enhances data quality by ensuring the dataset is accurate and representative, thus preventing misleading conclusions. It also improves model performance by reducing noise and mitigating overfitting.

The sequence of predicted values can be noisy, especially in the 1.5-second windows, so a median filter was applied to smooth the data and reduce the impact of outliers. The median filter works by sliding a window of a specified size over the dataset (in this case, the window size was 3), replacing each data point with the median value from the values within the window. The illustration in the picture 4.11 depicts the functionality of the medium filter.

The alternative approach employed a specialised filter with the All-in-One segments. Due to the varying segment sizes, using a median filter was unsuitable. Therefore, a filter was designed to evaluate the length of each segment. For segments shorter than one second, the filter checked if the anterior and posterior segments shared the same quadrant. If they did, the filter updated the quadrant value of the current segment to match the quadrant values of the previous and subsequent segments. This meticulous methodology ensured seamless continuity and coherence, enabling the filter to efficiently process the data while maintaining precision and integrity. The picture 4.12 provides a visual breakdown of how the filter functions.

Evaluation

The model's performance was evaluated using the same metrics outlined in Section 2.6. From all computed metrics, F1-score and the confusion matrix are presented in the next section, since both provide a comprehensive understanding of the model's classification abilities.

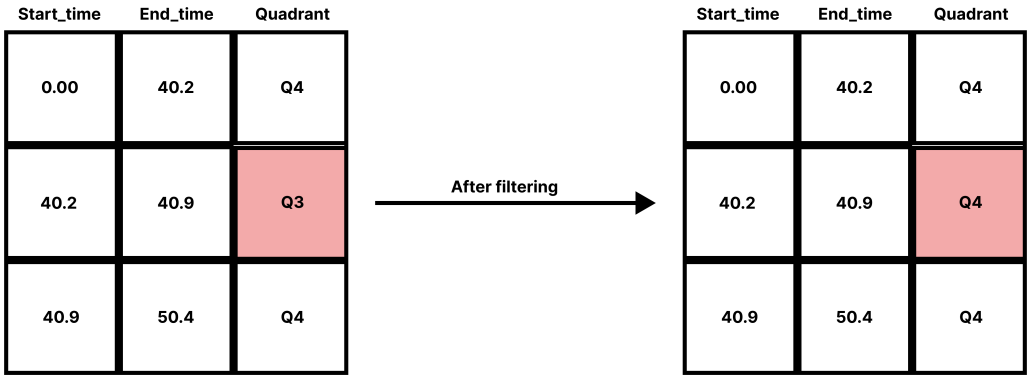


Figure 4.12: Concept of custom filtering in data processing.

The first metric is the average F1-score, calculated for all folds, repetitions, and quadrants. As previously discussed in Section 2.6, the F1-score is used because it balances precision and recall, offering a single metric that reflects both false positives and false negatives. Standard deviation calculations for the overall F1-score and each quadrant also help gauge the variability and consistency of the model's performance across different segments.

The second key metric is the percentage-based confusion matrix. This matrix provides a detailed view of the model's classification accuracy by showing the percentage of correctly classified values for each combination of true and predicted classes calculated across all folds. Similar to the F1-score, we compute the average percentage and the standard deviation for each cell of the confusion matrix to evaluate the model's reliability and the distribution of misclassifications.

4.3.3 Results and Discussion

In this section, we present and analyze the results of our experiments following classical methodologies. We conducted the experiments using a 3-fold cross-validation approach with 30 repetitions (30x3-fold CV) and a comprehensive evaluation of two distinct datasets. Tables 4.5 and 4.6 present a comprehensive summary of the results for the 30x3-fold CV experiment and the experiment using the two datasets, respectively.

	1.5 standard	All-in-One standard	1.5 novel	All-in-One novel
Q1	68.90%	36.30%	67.80%	37.10%
Q2	62.40%	24.50%	62.90%	27.30%
Q3	24.60%	19.50%	25.40%	19.60%
Q4	51.20%	26.70%	53.90%	28.40%
Weighted Avg	**55.10%**	**29.90%**	**55.90%**	**30.60%**

Table 4.5: F1-score obtained for the 30x3-fold CV experiment using only the 34-song dataset per quadrant.

	1.5 standard	All-in-One standard	1.5 novel	All-in-One novel
Q1	57.74%	56.98%	57.77%	56.25%
Q2	46.62%	50.46%	49.06%	44.65%
Q3	42.23%	48.00%	40.94%	44.24%
Q4	59.98%	54.72%	60.06%	50.00%
Weighted Avg	**52.97%**	**53.17%**	**53.38%**	**49.55%**

Table 4.6: F1-score obtained with the static MER and MEVD dataset experiment per quadrant.

Experiments		Confusion Matrix (in percentage)			
		Q1	Q2	Q3	Q4
1.5 Standard	Q1	**57.65%**	27.60%	7.38%	7.38%
	Q2	36.22%	**42.56%**	12.95%	8.26%
	Q3	3.66%	3.85%	**46.15%**	46.34%
	Q4	10.34%	3.41%	25.06%	**61.19%**
All-in-One Standard	Q1	**60.77%**	23.18%	7.27%	8.78%
	Q2	30.66%	**48.44%**	14.08%	6.82%
	Q3	5.92%	0.91%	**51.14%**	42.03%
	Q4	20.58%	2.98%	18.14%	**58.30%**
1.5 Novel	Q1	**57.10%**	21.58%	14.81%	6.50%
	Q2	34.57%	**43.03%**	16.26%	6.15%
	Q3	3.85%	5.25%	**50.56%**	40.34%
	Q4	10.17%	2.22%	29.03%	**58.58%**
All-in-One Novel	Q1	**65.27%**	17.05%	7.76%	9.92%
	Q2	40.26%	**39.86%**	8.83%	11.05%
	Q3	6.50%	0.91%	**50.93%**	41.66%
	Q4	18.43%	5.25%	19.95%	**56.37%**

Table 4.7: Confusion Matrix for the static MER and MEVD dataset experiment (in percentage).

In our 30x3-fold CV experiment, we observed that the 1.5-second segments achieved F1-scores of 55.10% with standard features and 55.90% with novel features, indicating that the novel features slightly outperformed the standard ones. However, the results suggest room for improvement and that the small dataset size may have affected the lower classification performance.

Conversely, when employing the All-in-One approach, we observed significantly lower scores of 29.90% and 30.60% for standard and novel features, respectively.

The statistical tests confirm these observations:

1. For 1.5-second segments, the comparison between novel and standard features showed no significant difference (p-value = 0.50061), indicating that while novel features performed slightly better, the improvement was not statistically significant.

2. Similarly, for the All-in-One approach, the comparison between novel and standard features also showed no significant difference (p-value = 0.55895).

3. However, when comparing the 1.5-second segments to the All-in-One approach, both for novel features (p-value $= 7.1982 \times 10^{-55}$) and standard features (p-value $= 2.5008 \times 10^{-51}$), the differences were statistically significant. This indicates a clear difference in performance based on the segmentation method used.

Interestingly, when using novel features for All-in-One segments, the top-ranked features were mainly standard features. Only a few new features made it to the top, justifying the small improvement from standard to novel features and suggesting that most novel features do not improve the outcomes for All-in-One segments. The increased complexity of new features comes from their specificity in capturing a single emotion. If a segment contains multiple emotions, these features provide little to no valuable information, resulting in poorer performance. Essentially, these features are more complex than standard ones and are effective only when a segment has a single emotional tone. When multiple emotions are present, the effectiveness of these new features decreases significantly.

The lower results could be attributed to the small dataset size, which may have contributed to challenges in the classification process. Moreover, the obtained segments may each contain multiple emotions. The absence of emotional consistency within the segments makes it challenging to classify them, possibly contributing to the low score achieved. Figure 4.13 illustrates the predicted and real emotional variation for one song.

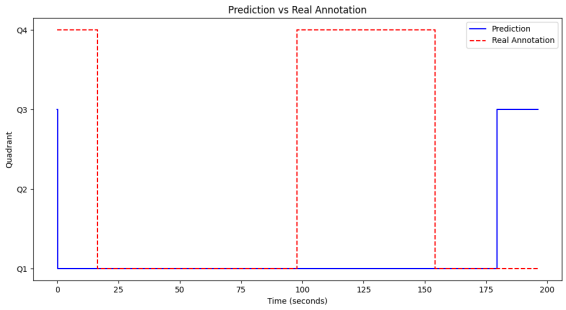


Figure 4.13: Comparison between the annotated and predicted emotion quadrants for the song "Whenever, Wherever" by Shakira using the All-in-One segmentation approach.

In this particular song, there exists a segment that spans from 97 seconds to 124 seconds. According to the annotation, the segment is divided into two parts: from 97 seconds to 100 seconds, characterized as Q1, and from 100 seconds to 124 seconds, characterized as Q4. This suggests that the segment in question may potentially encompass multiple emotional shifts.

As for our other experiment, we trained our model using the MERGE Audio Complete dataset and tested it on the MEVD dataset. When using 1.5-second

segments, we saw the F1-score increase from 52.97% to 53.17% with the addition of new features. Furthermore, the All-in-One approach slightly improved over the 1.5-second segments, increasing the F1-score from 52.97% to 53.17% when using standard features.

It is important to note that the variable windows produced by the All-in-One segmentation method yielded better results than the fixed 1.5-second windows for the third quadrant. This quadrant typically yields the worst results because it is challenging for models to recognize, likely because it also takes the longest for people to identify (1446 ms sad (Q3), [Vieillard et al., 2008]). Additionally, the table 4.7 shows that the third quadrant is often mistaken for the fourth. This is likely due to the low energy levels associated with both quadrants, making it difficult for the model to distinguish between the subtle emotional differences.

However, the novel features did not perform as well as the standard features when using the All-in-One approach, decreasing from 53.17% to 49.55%. The decrease in performance for the All-in-One segments is attributed to the fact that feature ranking was computed for the MERGE Audio Complete dataset, not the MEVD dataset, causing many novel features to rank highly. As seen in the 30x3-fold CV cross-validation experiment, the All-in-One segments performed poorly with most novel features. These features are more intricate than standard ones and are only adequate for segments with a single emotional tone. All-in-One segments often contain multiple emotions, as shown in Figure 4.13, causing top features to offer little information, resulting in poorer outcomes.

4.4 DL Approach

The upcoming section explores the application of deep learning techniques to improve MEVD, with a particular focus on the learning capabilities of CNNs. Drawing from the groundwork laid by [Louro et al., 2024a], this section delves into the integration of segmentation models within the framework of music analysis, exploring the impact of segmentation tools on feature extraction from song excerpts. It aims to assess their efficacy in enhancing music analysis outcomes by comparing the results obtained with and without their utilization, akin to the previous section.

4.4.1 Datasets

The dataset used to train and test the CNN model remained consistent with the specifications in Section 4.3. However, while the original dataset consisted of 3,554 clips, each 30 seconds long, for this experiment, each of these clips was divided into 1.5-second segments. This ensured that the segments used in the training and testing phases had the same length, aligning the experiment to maintain consistent segment durations.

Given the model's unsatisfactory performance with the MERGE Audio Complete dataset, we suspected this might be due to its unbalanced nature. To test this

hypothesis, we utilized the MERGE Audio Balanced dataset, described in Section 3.1, to investigate if a balanced dataset could enhance the model's performance. This balanced dataset was specifically chosen for training in this experiment.

4.4.2 Methodology

This methodology section proposes an approach that employs CNN-based architectures to classify segments into one of the four quadrants of Russell's Circumplex model. Besides experimenting with the CNN-based architecture, a simple Dense Neural Network (DNN) was also experimented with the optimal feature set found in Section 4.3. Figure 4.14 gives an overview of the complete process. The goal is to compare the standard approach using a 1.5-second window and the All-in-One approach using the segment windows produced by the All-in-One tool and find out if segmentation tools are beneficial and contribute to better results.

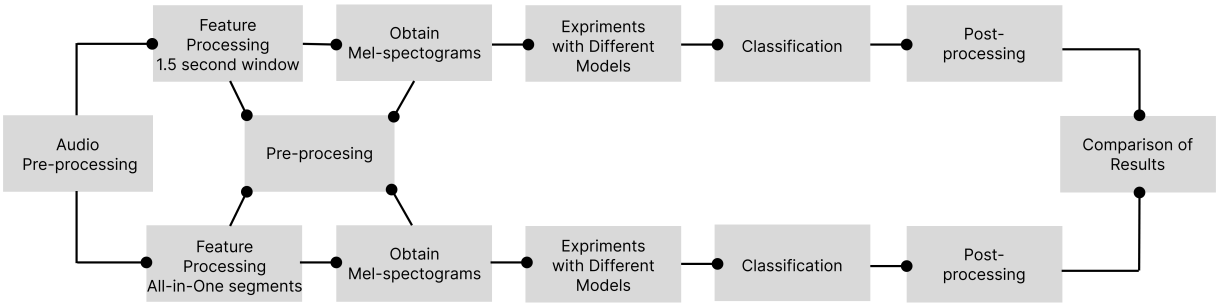


Figure 4.14: High level overview of the methodology.

Pre-processing

This section will discuss the pre-processing steps required to obtain the features and Mel-spectrograms needed for our CNN and DNN models, respectively.

Feature processing encompasses feature extraction, dimensionality reduction, and feature selection, all conducted within the same workflow. These steps are summarized here, as previously detailed in Section 4.3. First, feature extraction was performed on the static 1.5-second windows and the dynamic windows generated by the All-in-One segmentation tool. After extraction, dimensionality reduction was applied to remove features with zero standard deviation and reduce multicollinearity by eliminating highly correlated features, as previously discussed. Finally, feature selection was done using the RelieF algorithm, which ranks and selects the most critical features for classification.

To obtain the Mel-spectrograms, the first step is processing the audio files, which are loaded and converted into a standard WAV format. The audio files were then normalized and downsampled to a sampling rate of 16,000 Hz. Mel-spectrograms were generated for each audio file by transforming the audio signals into the frequency domain. This transformation involved applying a filter bank to extract

frequency bands and computing the spectral power's logarithm to obtain Mel-scale magnitudes. Padding was added to ensure uniform dimensions across all spectrograms.

For the approach involving 1.5-second length clips, each Mel-spectrogram captured the acoustic features of a 1.5-second segment extracted from the respective music pieces. In contrast, with the All-in-One approach, each Mel-spectrogram represented the aggregate acoustic characteristics of entire segments within the music compositions, such as verses, choruses, or bridges, which encapsulate distinct musical elements and structural components. However, since these segments can vary in length, and CNN requires a fixed-length input, each segment was divided into mini-segments of 1.5 seconds. Padding was applied to ensure consistency across all mini-segments. The last mini-segment was used when necessary, guaranteeing that all inputs fed into CNN were the same length.

Overview of the Model Architectures

We conducted two primary experiments using standalone deep learning-based architectures for MEVD. The first experiment used Mel-spectrograms, while the second utilized pre-extracted novel features from the audio signals.

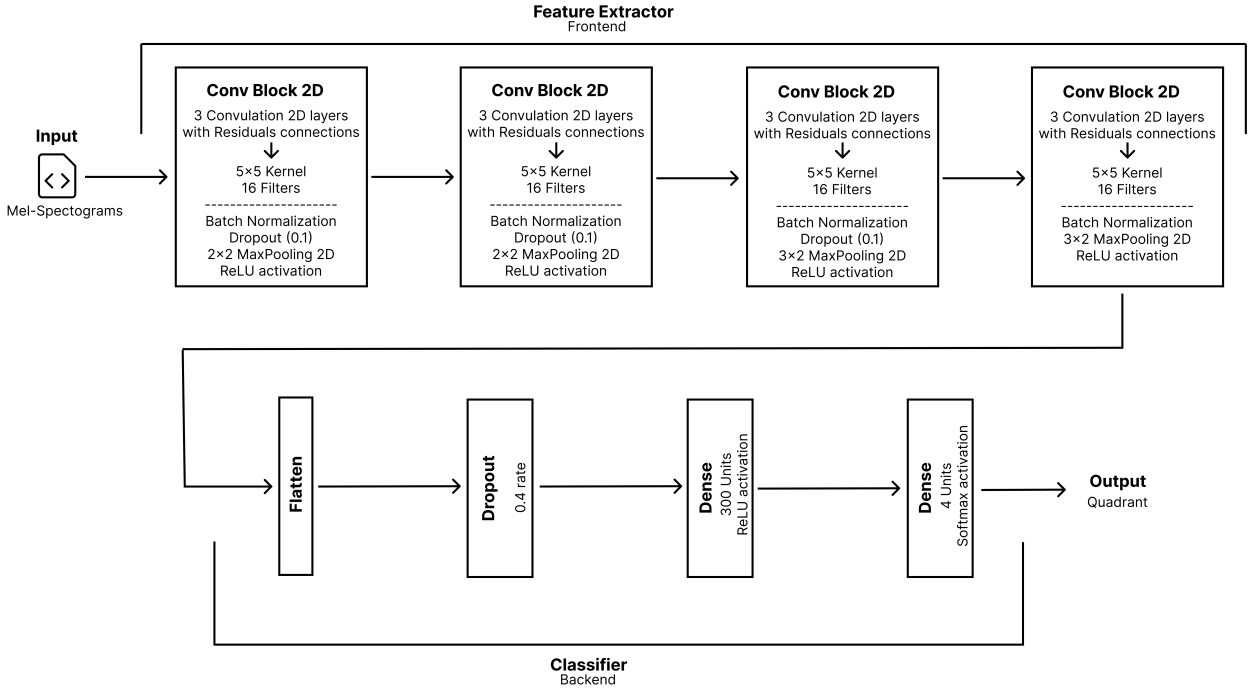


Figure 4.15: Architeture of the first CNN model.

For the first experiment, the model, as seen in Figure 4.15 includes four convolutional layers followed by max pooling, batch normalization, which comprise the feature learning portion. The classification portion includes a dropout layer, followed by two fully connected layers. Each convolutional layer features 16 filters with a (3, 3) kernel size and ReLU activation, while max pooling layers reduce

spatial dimensions. Batch normalization stabilizes activations, and dropout prevents overfitting.

Fully connected layers end with softmax activation for multi-class classification, in our case outputting one of the four quadrants of Russell's Circumplex model. Optimization is conducted with the stochastic gradient descent optimizer which the objective of minimizing categorical cross-entropy loss. Designed for processing Mel-spectrograms.

We also experimented with a CNN and LSTM architecture, designed to better capture time-related features. The critical modification involved configuring the CNN layers to handle time series inputs, allowing the model to process each time step independently. By introducing an LSTM layer after the convolutional and pooling layers, the model became better equipped to learn temporal dependencies and capture the dynamic nature of emotional variability. This LSTM layer, whether configured as unidirectional to utilize past information or bidirectional to leverage both past and future data, enhances the classification of sequences based on temporal patterns.

These adjustments were intended to enable the CNN-based model to effectively process and understand audio sequences, capturing temporal dependencies, enhancing feature learning, and improving performance in emotion variation detection. The CNN layers independently extract features at each time step. At the same time, the LSTM component ensures that the model comprehends the flow and changes in emotion over time, resulting in more nuanced and accurate predictions.

On the other hand, the model used for the second experiment was a DNN model. Instead of learning features directly from the data, this model takes an array of pre-extracted features (in this case, the standard plus novel features employed in the classical approach) and feeds them into the network. In this DNN approach, the network processes these manually extracted features, selecting the most relevant ones for classification. This method allows the model to focus on feature selection and classification rather than feature extraction. Figure 4.16 shows a visual representation of the model.

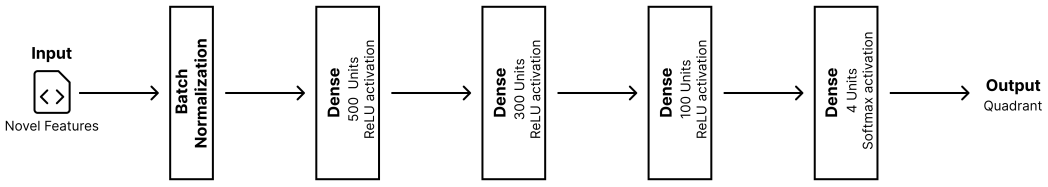


Figure 4.16: Architeture of the DNN model.

Upon flattening the output, a dropout layer with a 40% rate is added to prevent overfitting further. Subsequently, the model includes a dense layer with 300 neurons and ReLU activation, followed by a final dense layer with four neurons and softmax activation to obtain quadrant probabilities.

Training Phase

During the training phase, we optimized the model using the training dataset and conducted hyperparameter tuning with Keras Tuner [O'Malley et al., 2019]. This process involved systematically exploring various combinations of hyperparameters, such as batch sizes, optimizers, and corresponding learning rates. Each configuration was evaluated on a validation dataset, and we used Bayesian optimization to refine the hyperparameters iteratively. This approach aimed at maximizing accuracy based on the validation set.

To prevent overfitting, we stopped training once the model reached the accuracy threshold. After conducting ten trials, we stored the results in a folder. The trial that yielded the highest model performance was selected, and its corresponding hyperparameters were retained for use in the testing phase.

Segment Prediction

As previously discussed in Section 4.3, after training, the computed hyperparameters are applied to the model, which is then used to predict the class of each fixed or variable segment. Figure 4.17 provides a comparison between the actual annotations and the model's predictions for the variable segments.

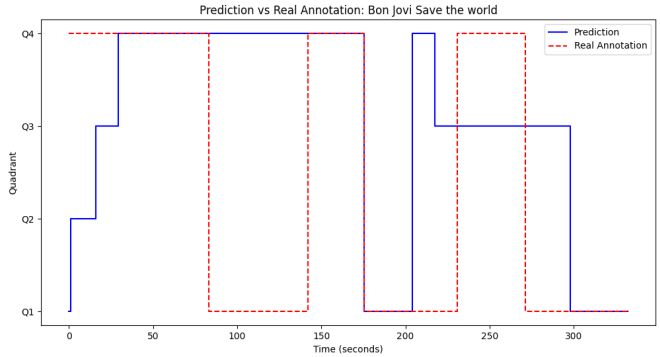


Figure 4.17: Comparison between ground truth annotations and model predictions for variable segments.

Post-Processing

As mentioned in the classical approach (Section 4.3), post-processing is essential to enhance data quality by identifying and removing potential outliers. This helps prevent misleading conclusions and improves model performance by reducing noise and mitigating overfitting. In the deep learning approach, similar post-processing techniques were employed. The 1.5-second-length segments were processed using a median filter to smooth data and reduce the impact of outliers, just as in the classical method. Additionally, a custom filter was applied for the All-in-One segments with varying sizes to ensure continuity and coherence, effectively handling segments shorter than one second by aligning them with adjacent segments where appropriate. These post-processing steps were crucial in

refining the output of the deep learning models, ensuring that the predictions were robust and reliable.

Evaluation

The same evaluation metrics outlined in Section 4.3.2 were applied here for the DL approach. These include the average F1-score, which balances precision and recall, and the percentage-based confusion matrix, as referenced earlier. These metrics thoroughly understand the model's performance, including its variability and consistency across folds and quadrants.

4.4.3 Results and Discussion

In this chapter, the outcomes of the experiments are discussed and analyzed. The research focused on using Mel-spectrograms along with additional features. Tables 4.8 and 4.12 present a detailed summary of the findings, with Table 4.8 showcasing the analysis using Mel-spectrogram and Table 4.12 covering the results using features.

F1-score and confusion matrix for our mel-spectogram experiments are depicted in Table 4.8 and Table 4.9. It is possible to observe that the standard CNN model and the LSTM model using fixed 1.5 second window achieve an F1 Score of 20.61% and 17.14%, respectively, and the standard CNN outperformed the LSTM model. However, by examining the table, we can see that the fixed window approach significantly improves the third quadrant, with the LSTM model increasing from 5.11% to 11.77%. This improvement suggests that utilizing an LSTM model may be necessary to capture temporal context effectively, which is crucial for improving results for the third quadrant, typically the most challenging segment in MEVD.

	CNN		CNN LSTM	
Quadrant	1.5	All-in-One	1.5	All-in-One
Q1	46.42%	44.19%	7.33%	0.00%
Q2	6.83%	22.17%	41.17%	39.59%
Q3	5.11%	0.00%	11.77%	3.94%
Q4	14.71%	0.00%	9.84%	1.44%
Weighted Avg	**20.61%**	**18.37%**	**17.14%**	**10.91%**

Table 4.8: F1-score obtained in MERGE Audio Complete for the Mel-spectrograms experiment using different models per quadrant.

One particularly concerning observation is the presence of 0.00% F1-scores in certain quadrants when using the segments generated for the All-in-One approach. This approach divided the original segments into 1.5-second mini-segments because the training was conducted using 1.5-second Mel-spectrograms. Consequently, during testing, the segments had to be broken down into these 1.5-second mini-segments, with each one fed individually to the model. The final

Experiments		Confusion matrix (in percentage)			
		Q1	Q2	Q3	Q4
CNN 1.5	Q1	**30.85%**	25.53%	16.24%	27.36%
	Q2	32.62%	**39.71%**	12.76%	14.89%
	Q3	39.23%	5.38%	**23.07%**	32.30%
	Q4	6.22%	8.65%	33.91%	**51.21%**
CNN All-in-One	Q1	**30.59%**	24.30%	18.57%	26.52%
	Q2	25.47%	**23.04%**	19.50%	31.97%
	Q3	11.18%	78.21%	**0.00%**	10.60%
	Q4	0.00%	0.00%	0.00%	**0.00%**
CNN LSTM 1.5	Q1	**61.20%**	14.65%	5.17%	18.96%
	Q2	30.40%	**26.11%**	16.33%	27.14%
	Q3	19.15%	3.73%	**34.57%**	42.52%
	Q4	5.35%	8.92%	30.35%	**55.35%**
CNN LSTM All-in-One	Q1	**0.00%**	0.00%	0.00%	0.00%
	Q2	29.23%	**25.04%**	19.04%	26.66%
	Q3	29.37%	22.04%	**7.82%**	40.74%
	Q4	18.15%	2.79%	48.45%	**30.59%**

Table 4.9: Confusion Matrix for MERGE Audio Complete for the Mel-spectrograms experiment using different models (in percentage).

prediction for the entire segment was then determined by taking the mode of the predicted quadrants across the mini-segments of the segment. For example, suppose the actual quadrant for a segment is Q2, but the model consistently predicts Q3 for the mini-segments. In that case, it severely diminishes the F1 Score for that quadrant, potentially resulting in an F1-score of 0.00%.

Another factor contributing to the poor performance could be the training approach, which, although it uses 30-second clips, divides them into 1.5-second segments. This segmentation might limit the model's ability to capture the songs' broader temporal structures and emotional nuances, as each 30-second clip typically contains limited emotional variation. Additionally, the problem arises because the training was done using 30-second clips instead of full songs. This approach prevents the model from capturing the whole emotional progression and structural segmentation that would be present in a complete song, leading to suboptimal generalization across different quadrants, as the model misses out on understanding the broader emotional context within a song.

	CNN 1.5	CNN All-in-One
Q1	8.90%	42.53%
Q2	41.05%	36.20%
Q3	7.55%	0.00%
Q4	10.57%	0.00%
Weighted Avg	**17.06%**	**21.35%**

Table 4.10: F1-score obtained in the MERGE Audio Balanced dataset for the Mel-spectograms experiment per quadrant.

Due to the low results shown in Table 4.8, further experiments were conducted with the MERGE Audio Balanced dataset shown in Figure 3.2, a balanced dataset across the quadrants to determine if the issue was due to the unbalanced nature of the training dataset. The F1-score and confusion matrix results for this experiments are presented in Table 4.10 and Table 4.11.

Experiments		Confusion Matrix (in percentage)			
		Q1	Q2	Q3	Q4
CNN 1.5	Q1	**69.09%**	26.39%	0.69%	3.80%
	Q2	30.44%	**26.02%**	16.35%	27.17%
	Q3	71.00%	0.29%	**13.01%**	15.68%
	Q4	2.11%	4.23%	40.21%	**53.43%**
CNN All-in-One	Q1	**30.11%**	21.15%	21.35%	27.37%
	Q2	26.96%	**33.04%**	11.97%	28.01%
	Q3	0.00%	0.00%	**0.00%**	0.00%
	Q4	0.00%	0.00%	0.00%	**0.00%**

Table 4.11: Confusion Matrix for MERGE Audio Balanced dataset for the Mel-spectograms experiment (in percentage).

The results in Table 4.10 reveal an increase in the F1 score when using the All-in-One windows, compared to the experiment conducted using the MERGE Audio Complete Dataset for training. Using variable windows, combined with training on a balanced dataset, may help improve the results. However, further experiments were not conducted due to time constraints and the high computational cost involved.

Shifting our focus to the experiments involving features as input to DNN architectures, the F1-score and confusion matrix results are presented in Table 4.12 and Table 4.13.

	Standard Features		Novel Features	
Quadrant	DNN 1.5	DNN All-in-One	DNN 1.5	DNN All-in-One
Q1	47.18%	33.40%	38.77%	36.80%
Q2	40.44%	30.49%	8.87%	9.76%
Q3	40.06%	41.95%	22.68%	23.27%
Q4	47.96%	38.55%	35.08%	42.28%
Weighted Avg	**44.52%**	**35.58%**	**27.60%**	**29.00%**

Table 4.12: F1-scores from the features experiment using the MERGE Audio Complete dataset, evaluated across each quadrant.

Interestingly, as shown in Table 4.6 using SVM with features, the DNN model also demonstrated improved performance in the third quadrant when using the All-in-One segments compared to the 1.5-second segments.

The DNN model with the All-in-One segments also showed better results for the standard features, similar to what was observed in the SVM experiment.

Experiments		Confusion Matrix (in percentage)			
		Q1	Q2	Q3	Q4
DNN 1.5 Standard	Q1	**45.62%**	31.58%	8.19%	14.59%
	Q2	36.15%	**37.07%**	13.15%	13.61%
	Q3	8.53%	4.22%	**47.56%**	39.68%
	Q4	13.40%	4.37%	34.60%	**47.61%**
DNN All-in-One Standard	Q1	**45.09%**	37.12%	4.91%	12.86%
	Q2	30.42%	**42.75%**	8.77%	18.04%
	Q3	23.14%	18.69%	**29.84%**	28.31%
	Q4	29.76%	18.92%	12.15%	**39.15%**
DNN 1.5 Novel	Q1	**39.60%**	32.47%	7.12%	20.79%
	Q2	22.94%	**13.49%**	34.95%	28.60%
	Q3	26.51%	26.91%	**17.43%**	29.13%
	Q4	25.99%	18.29%	19.893%	**35.82%**
DNN All-in-One Novel	Q1	**40.30%**	41.80%	2.13%	15.7%
	Q2	46.56%	**42.41%**	0.00%	11.02%
	Q3	37.57%	18.68%	**17.46%**	26.27%
	Q4	13.09%	19.00%	30.89%	**37.00%**

Table 4.13: Confusion Matrix for the features experiment using the MERGE Audio Complete dataset (in percentage).

Additionally, in the MEVD dataset, a 3-fold with 30 repetitions (30x3-fold CV) was conducted for the DNN model, following a similar approach used for the CNN. This was carried out to determine whether using full-length songs would allow the DNN model to capture better and generalize emotional variations compared to using the 30-second clips from the MERGE Audio Complete dataset. The F1-score and confusion matrix results are shown in Table 4.14 and Table 4.15.

	Standard Features		Novel Features	
Quadrant	1.5 DNN	All-in-One DNN	1.5 DNN	All-in-One DNN
Q1	65.30% ±0.100	34.81% ±0.047	65.00% ±0.112	24.75% ±0.057
Q2	58.70% ±0.146	12.48% ±0.050	58.80% ±0.134	43.85% ±0.092
Q3	19.40% ±0.097	24.12% ±0.026	20.30% ±0.099	2.85% ±0.051
Q4	46.70% ±0.089	46.01% ±0.023	46.80% ±0.099	21.71% ±0.085
Weighted Avg	**50.80%**	**31.09%**	**51.20%**	**24.38%**

Table 4.14: F1-scores from the features experiment on the MEVD dataset using 30x3-fold CV, comparing standard and novel feature sets across each quadrant.

The results in Table 4.14 show that, although the 1.5-second segments achieved a higher overall weighted F1-score compared to the All-in-One approach, this improvement is mainly due to the larger number of samples available in the 1.5-

Experiments		Confusion Matrix (in percentage)			
		Q1	Q2	Q3	Q4
DNN 1.5 Standard	Q1	**64.61%**	18.19%	4.44%	12.74%
	Q2	20.37%	**64.07%**	6.30%	9.24%
	Q3	9.79%	12.31%	**22.30%**	55.58%
	Q4	11.70%	9.72%	34.38%	**44.19%**
DNN All-in-One Standard	Q1	**39.74%**	34.65%	9.76%	15.82%
	Q2	34.60%	**16.09%**	13.93%	35.37%
	Q3	28.89%	28.40%	**20.88%**	21.81%
	Q4	19.57%	12.57%	26.41%	**41.44%**
DNN 1.5 Novel	Q1	**66.57%**	17.79%	4.30%	11.32%
	Q2	20.91%	**62.27%**	7.18%	9.62%
	Q3	10.24%	12.33%	**22.58%**	54.83%
	Q4	11.66%	9.27%	33.90%	**45.15%**
DNN All-in-One Novel	Q1	**53.74%**	7.54%	7.88%	30.82%
	Q2	20.84%	**29.48%**	20.84%	28.83%
	Q3	30.48%	28.13%	**6.33%**	35.03%
	Q4	39.92%	8.79%	19.75%	**31.52%**

Table 4.15: Confusion Matrix for the features experiment on the MEVD dataset using 30x3-fold CV (in percentage).

second segments.

Moreover, consistent with the results of the last experiment with the MERGE Audio Complete dataset, Table 4.12, the standard features yielded better results than the novel features in All-in-One segmentation approaches.

4.5 Hybrid Approach

The upcoming section explores using a hybrid deep learning model to improve MEVD by combining the strengths of CNN and DNN architectures. Building on Chapter 4.4, this hybrid model integrates CNNs to capture spatial features from Mel-spectrograms and DNNs to process traditional features. The goal is to evaluate whether this combined approach enhances emotion detection in music compared to standalone models.

Dataset

The datasets used in this section are the same as those utilized in Chapter 4.3. Therefore, this study primarily relied on the MERGE Audio Complete dataset for training and in the MEVD dataset for testing.

Methodology

The methodology employed in this chapter is consistent with the approach detailed in Chapter 4.4.2, except for the model architecture used.

In this chapter, a hybrid model is introduced to integrate the strengths of both CNN and DNN architectures, providing a more comprehensive approach for MEVD. After experimenting with two separate models, one utilizing Mel-spectrograms with a CNN architecture and the other employing pre-extracted features with a DNN, we adapted a third approach that combines both strengths. This hybrid architecture, depicted in Figure 4.18, merges the CNN and DNN branches.

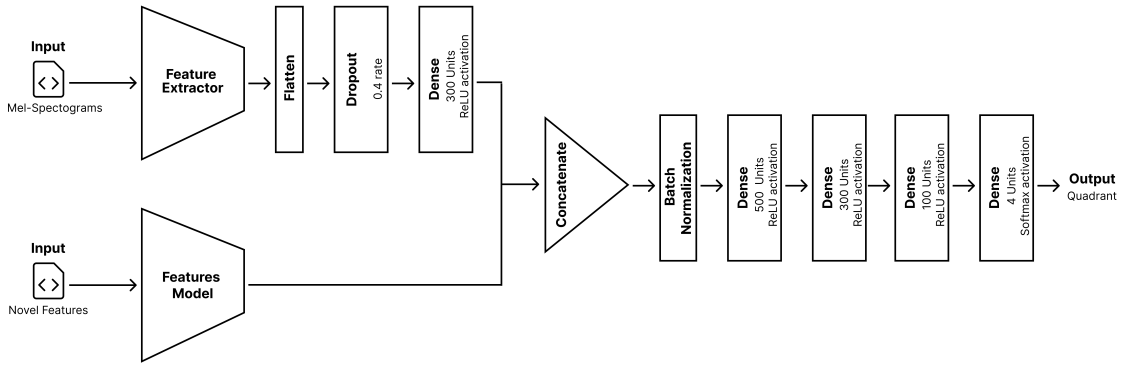


Figure 4.18: Architeture of the hybrid model.

In constructing this ensemble model, the DNN portion was pre-trained separately to identify the best-performing model. The optimal weights from this training phase were saved, and when training the entire hybrid model, these weights were loaded, and the DNN layers were frozen. This approach ensured that the DNN branch retained its fine-tuned parameters while the CNN branch could further adapt and learn from the Mel-spectrogram inputs. These two branches work together, with the final output layer combining the processed information to classify the audio sample into one of the four quadrants of Russell's A/V model.

Since we did not have the features extracted for 1.5-second windows from the MERGE Audio Complete dataset and extracting them would not be feasible in terms of time, we adopted an alternative approach. We passed the features of the 30-second excerpts to each corresponding 1.5-second window within that 30-second clip. This method aimed to enable the model to learn both global (30-second) and local (1.5-second) features, respectively, leveraging broader context while focusing on finer details for more accurate classification.

Exclusively for training the CNN portion, we utilized classic data augmentation methods, as described in Section 2.5, to increase the training data. This approach aimed to enhance the model's performance, as neural networks benefit from large amounts of data.

Results and Discussion

The F1-score and confusion matrix results of the hybrid model evaluation are shown in Table 4.16 and Table 4.17.

Quadrant	Standard Features		Novel Features	
	1.5 Hybrid	All-in-One Hybrid	1.5 Hybrid	All-in-One Hybrid
Q1	37.35%	40.96%	31.61%	29.53%
Q2	42.13%	48.70%	32.65%	23.75%
Q3	26.70%	25.43%	17.15%	27.64%
Q4	48.51%	32.69%	53.01%	43.96%
Weighted Avg	**39.86%**	**36.94%**	**35.45%**	**31.22%**

Table 4.16: F1-score obtained in the MERGE Audio Complete with the hybrid model per quadrant.

Experiments		Confusion Matrix (in percentage)			
		Q1	Q2	Q3	Q4
Hybrid 1.5 Standard	Q1	**42.95%**	24.66%	13.50%	18.87%
	Q2	40.49%	**41.13%**	11.88%	6.48%
	Q3	17.26%	21.38%	**23.58%**	37.76%
	Q4	19.78%	12.29%	19.78%	**48.14%**
Hybrid All-in-One Standard	Q1	**39.85%**	17.96%	21.26%	20.90%
	Q2	23.64%	**52.38%**	12.83%	11.13%
	Q3	26.22%	13.87%	**25.48%**	34.41%
	Q4	10.41%	20.11%	37.99%	**31.48%**
Hybrid 1.5 Novel	Q1	**36.52%**	32.70%	13.76%	17.00%
	Q2	47.49%	**35.21%**	9.72%	7.56%
	Q3	23.54%	27.91%	**18.60%**	29.93%
	Q4	19.42%	13.52%	22.33%	**44.71%**
Hybrid All-in-One Novel	Q1	**34.93%**	26.38%	29.04%	9.64%
	Q2	26.97%	**56.47%**	6.27%	10.27%
	Q3	25.43%	22.80%	**27.54%**	24.21%
	Q4	20.87%	21.40%	24.73%	**32.99%**

Table 4.17: Confusion Matrix for the MERGE Audio Complete with the hybrid model (in percentage).

The results from Table 4.16 indicate that the All-in-One segmentation did not outperform the 1.5-second segments. A consistent finding throughout these experiments was that, for the variable windows produced by the All-in-One approach, novel features did not improve the results compared to standard features. This may be because many novel features rank highly, as feature ranking was based on the MERGE Audio Complete dataset rather than the MEVD dataset. As previously stated in Section 4.3, the All-in-One segments performed worse with novel features because these features are more complex and require a single emotional tone. In contrast, the variable segments often contain multiple emotions.

A 30x3-fold CV experiment was planned using the MEVD dataset for the hybrid model. Similar to the approach used for the CNN, it was intended to divide the All-in-One segments into 1.5-second mini-segments to maintain consistency between training and testing sets. This strategy aimed to explore whether using the MEVD dataset, with its full-length songs, could better capture and generalize emotional variations compared to the 30-second clips in the MERGE Audio Complete dataset. The objective was to enhance model performance by exposing the network to more comprehensive emotional shifts and patterns within the music. However, due to time constraints, this experiment could not be conducted.

4.6 Summary

In Chapter 4, the research applies the previously discussed methodologies to a series of experiments. The chapter starts by replicating Panda and Paiva's work on emotion tracking using SVMs and audio features, revealing challenges in achieving comparable results despite translating the MATLAB code to Python. This highlights the importance of ensuring research replicability across different platforms to validate findings. The replication experiments, divided into three parts using MIR Toolbox, Marsyas, and their combination, involved SVC and SVRs for quadrant, arousal, and valence prediction, showing the necessity of tailored hyperparameter tuning for different feature types. The chapter then evaluates the effectiveness of DeepChorus and All-in-One segmentation tools across various music genres as alternatives to fixed 1.5-second intervals. DeepChorus performed well in Country and Hip-Hop but struggled with Latin and Electronic due to their unconventional structures. Similarly, All-in-One showed potential in some genres but had difficulties with complex structural elements in others.

Next, the experiments were carried out using two distinct paradigms: classical and DL. The classical approach involved using SVMs alongside the All-in-One segmentation tool to see how traditional methods would perform when improved by advanced segmentation techniques. The use of these segmentation tools showed an improvement in the results. For example, experiments using standard features (as shown in Table 4.6) showed improved performance, particularly in the second and third quadrants.

The classical experiments revealed valuable model performance insights across different datasets and segmentation approaches. The 30x3-fold CV showed that novel features slightly outperformed standard ones with 1.5-second segments, though small dataset size likely hindered overall classification. While the All-in-One tool offered dynamic segment lengths, it struggled with novel features due to their complexity and multiple emotions within segments.

The static MERGE Audio Complete and MEVD dataset experiments highlighted the benefits of dynamic segment lengths for capturing nuanced emotions, particularly in the challenging second and third quadrants. However, the expected improvement from novel features did not materialize, emphasizing the need for more tailored feature ranking. These findings underscore the crucial role of dataset size and emotional consistency in achieving higher classification accuracy, with

the effectiveness of novel features depending on the data and segmentation approach used.

In contrast, the DL approach involved experimenting with various models, including a DNN, CNNs, a CNN combined with an LSTM layer, and a hybrid model that combines the CNN and DNN across different segmentation strategies.

When experimenting with Mel-spectrograms, as shown in Table 4.8, adding an LSTM layer to the CNN model improved performance in the third quadrant, which usually yields the worst results. This suggests that the LSTM layer, designed to capture temporal context, is beneficial for handling these more intricate emotional variations. However, it may require more sophisticated tuning or a more balanced dataset to fully realize its potential across all quadrants.

Due to the hypothesis that the unbalanced nature of the training dataset might be affecting the results, an experiment was conducted using the MERGE Audio Balanced dataset. Interestingly, when experimenting with the MERGE Audio Balanced dataset, the F1-score of the All-in-One approach improved, as seen in Table 4.10, suggesting that training with a balanced dataset may help enhance results for the variable window approach. Unfortunately, further experiments with the MERGE Audio Balanced dataset were not conducted due to time constraints and computational complexity.

The DNN approach, similar to the SVM results, demonstrated improved performance in the third quadrant when using All-in-One segments compared to 1.5-second segments, as shown in Table 4.12. The DNN also performed better with standard features, which was consistent with the SVM experiment. A 30x3-fold CV was conducted on the MEVD dataset to assess whether full-length songs would improve the model's ability to capture emotional variations compared to 30-second clips from the MERGE Audio Complete dataset. While the 1.5-second segments achieved a higher overall weighted F1-score, this was primarily due to the larger number of samples. Consistent with prior results, standard features outperformed novel features in the All-in-One segmentation approaches, as seen in Table 4.14.

The hybrid model, which combines the strengths of both CNN and DNN architectures, was evaluated to assess the potential of utilizing both Mel-spectrograms and traditional handcrafted features for emotion detection. As detailed in Table 4.16, the results demonstrated that the All-in-One segmentation tool did not outperform the 1.5-second segment.

Additionally, standard and novel features were tested across the SVM, DNN, and Hybrid models, with results consistently showing that standard features using the All-in-One segmentation yielded better outcomes. This can be attributed to the top-ranked features for All-in-One segments being mostly standard features, while novel features did not significantly impact. A possible explanation is that the increased complexity of novel features designed to capture specific emotions may be less effective when segments contain multiple emotions. If the All-in-One segments indeed encompass multiple emotional tones, novel features may struggle to provide meaningful information, leading to poorer performance.

As we move into Chapter 5, the final chapter of this work, we will consolidate the insights gained from the experiments and analyses conducted in Chapter 4. This next chapter will not only summarize the key findings but also provide a critical evaluation of the approaches used, discussing their strengths, limitations, and potential areas for improvement. Additionally, Chapter 5 will propose directions for future research, emphasizing how the methodologies and tools developed in this study can be refined and expanded to address the challenges of MEVD better.

Chapter 5

Conclusion and Future Work

This chapter summarizes the main contributions of this work, and proposes possible research directions for future work.

5.1 Conclusion

As this work comes to its end, the most relevant conclusions, limitations, and main contributions are presented and discussed.

One of the primary limitations affecting the methodologies' performance is the lack of emotionally relevant features that can be robustly generalised across different musical pieces. This issue is particularly challenging for the classical approach, where predefined features often fail to capture the diverse emotional nuances in music.

The most significant constraint for the deep learning approach is the need for larger amounts and better quality data to fully exploit these models' potential. As [Louro et al., 2024a] noted, the effectiveness of deep learning in MER is severely hampered by the lack of large, high-quality datasets, which are essential for these models to learn and generalise effectively.

Although the results were lower than expected, the All-in-One tool demonstrates potential, mainly when experimenting with a balanced dataset. The results showed performance improvement, suggesting that training with balanced data could bring out additional insights and patterns that needed to be fully highlighted in this study. This indicates that further exploration with balanced datasets could uncover more refined emotional variations and improve the overall effectiveness of the approach.

However, with further refinement, these tools hold significant promise for advancing the current state-of-the-art results in the MEVD field. Enhancements such as more effective feature selection, improved handling of variable window sizes, and deeper exploration of balanced datasets could significantly boost model accuracy and robustness. By addressing the challenges identified in this study—such as handling complex emotional segments and improving the distinction between

similar emotional quadrants—these tools can significantly improve emotion detection and classification in music, contributing to more accurate and nuanced models in the field.

Finally, the experiment that yielded the best results with the All-in-One segmentation approach was the hybrid architecture, which combines both features and mel-spectrograms. This suggests that future experiments should further utilise this architecture to explore its potential. By leveraging the complementary strengths of features and Mel-spectrograms, the hybrid approach offers a promising path forward, and refining this model could lead to even more significant improvements in emotion recognition and segmentation accuracy within the MEVD field.

5.2 Future Work

For future work, several promising directions remain open for exploration. A summary of these potential avenues is presented below:

- **Apply the Classic and DL approaches to the CAL500Exp dataset**, as this dataset may provide additional insights and potentially address some of the limitations encountered with smaller datasets.
- **Investigate Audio Large Language Models (LLMs) and embeddings** to enhance feature extraction and improve emotion recognition in music by leveraging advanced techniques in natural language processing and audio analysis.
- **Explore transformer architectures**, which support dynamic input sizes, enabling the use of All-in-One segmentation outputs without the need for padding to achieve uniform segment lengths, thereby taking full advantage of the structural segmentation to detect emotion variation.
- **Conduct experiments using the MERGE Audio Balanced dataset**, as this dataset offers an even distribution across emotional quadrants. This will allow for a more balanced and fair evaluation of the model's performance, potentially improving results by mitigating the impact of dataset imbalance observed in previous experiments.
- **Conduct future experiments using the hybrid model**, as this approach demonstrated the best results for variable window sizes in previous tests. Exploring further refinements and adjustments to this architecture could lead to significant improvements in emotion recognition.
- **Improve the All-in-One segmentation tool**, by refining its ability to capture and represent complex emotional shifts within music. Enhancing the tool's segmentation accuracy could address the limitations observed in this study and lead to better performance in Music Emotion Recognition systems.

References

Anna Aljanaki, Yi-Hsuan Yang, and M. Soleymani. Developing a benchmark for emotional analysis of music. *PLoS ONE*, 12, 2017. doi: 10.1371/journal.pone.0173392.

Thierry Bertin-Mahieux, Daniel Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, pages 591–596, 01 2011. URL https://ismir2011.ismir.net/papers/OS6-1.pdf.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA, 2006. ISBN 9780387310732. doi: 10.1117/1.2819119.

Rachel Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Bello. MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. *Proceedings - 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, 10 2014. URL https://rachelbittner.weebly.com/uploads/3/2/1/8/32182799/bittner_ismir_2014.pdf.

Jason Brownlee. *Probability for Machine Learning: Discover How to Harness Uncertainty with Python*. Course Hero, 2024. URL https://www.coursehero.com/file/90646292/Probability-for-Machine-Learning-Discover-How-To-Harness-Uncertainty-With-Python-by-Jason-Brownlee/. Accessed August 2024.

Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks, 2016. URL https://archives.ismir.net/ismir2016/paper/000009.pdf.

DatabaseCamp. Recurrent Neural Network - Types & Architecture, n.d. URL https://databasecamp.de/en/ml/recurrent-neural-network. Accessed: 2024-08-15.

Thomas Dixon. "Emotion": The History of a Keyword in Crisis. *Emotion Review*, 4(4):338–344, 2012. ISSN 1754-0739. doi: 10.1177/1754073912445814.

Yizhuo Dong, Xinyu Yang, Xi Zhao, and Juan Li. Bidirectional Convolutional Recurrent Sparse Network (BCRSN): An Efficient Model for Music Emotion Recognition. *IEEE Transactions on Multimedia*, 21(12):3150–3163, 2019. doi: 10.1109/TMM.2019.2918739.

Richard Duda, Peter Hart, and David G. Stork. *Pattern Classification*, volume xx. Wiley Interscience, 2001. ISBN 0-471-05669-3. URL https://www.researchgate.net/publication/228058014_Pattern_Classification. Revised edition.

Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4): 169–200, 1992. doi: 10.1080/02699939208411068.

Paul R. Farnsworth. A Study of the Hevner Adjective List. *The Journal of Aesthetics and Art Criticism*, 13(1):97–103, 1954. ISSN 00218529, 15406245. doi: 10.2307/427021.

Yazhong Feng, Yueting Zhuang, and Yunhe Pan. Popular music retrieval by detecting mood. In *Proc. 26th Int. ACM SIGIR Conf. on R&D in Information Retrieval*, pages 375–376, 07 2003. doi: 10.1145/860435.860508.

Alf Gabrielsson. Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, 5(1_suppl):123–147, 2001. doi: 10.1177/10298649020050S105.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. ISBN 9780262035613. URL http://www.deeplearningbook.org.

M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval*. ISMIR, 2002. URL https://ismir2002.ismir.net/proceedings/03-SP04-1.pdf.

Jacek Grekow. Music emotion recognition using recurrent neural networks and pretrained models. *Journal of Intelligent Information Systems*, 57:531 – 546, 2021. doi: 10.1007/s10844-021-00658-5.

Juan Sebastián Gómez Cañón, Estefanía Cano, Perfecto Herrera, and Emilia Gómez. Transfer learning from speech to music: towards language-sensitive emotion recognition models. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 136–140, 2021. doi: 10.23919/Eusipco47968.2020.9287548.

Qiqi He, Xiaoheng Sun, Yi Yu, and Wei Li. Deepchorus: A Hybrid Model of Multi-Scale Convolution And Self-Attention for Chorus Detection. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 411–415, 2022. doi: 10.1109/ICASSP43922.2022.9746919.

Kate Hevner. Experimental Studies of the Elements of Expression in Music. *The American Journal of Psychology*, 48(2):246–268, 1936. ISSN 00029556. doi: 10.2307/1415746.

Taejun Kim and Juhan Nam. All-In-One Metrical And Functional Structure Analysis With Neighborhood Attentions on Demixed Audio, 2023.

Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. A matlab toolbox for music information retrieval. In Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker, editors, *Data Analysis, Machine Learning and Applications*, pages 261–268. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. doi: 10.1007/978-3-540-78246-9_31.

Pedro Louro. MERGE Audio 2.0: Music Emotion Recognition next Generation - Audio Classification with Deep Learning. Master's thesis, University of Coimbra, 09 2022.

Pedro Lima Louro, Hugo Redinho, Ricardo Malheiro, Rui Pedro Paiva, and Renata Panda. A Comparison Study of Deep Learning Methodologies for Music Emotion Recognition. *Sensors*, 24(7), 2024a. ISSN 1424-8220. doi: 10.3390/s24072201.

Pedro Lima Louro, Hugo Redinho, Ricardo Santos, Ricardo Malheiro, Renato Panda, and Rui Pedro Paiva. MERGE – A Bimodal Dataset for Static Music Emotion Recognition, 2024b. URL https://arxiv.org/abs/2407.06060.

Ricardo Malheiro, Renato Panda, Paulo Gomes, and Rui Pedro Paiva. Bi-Modal Music Emotion Recognition: Novel Lyrical Features and Dataset. In *9th International Workshop on Music and Machine Learning – MML 2016 – in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – ECML/PKDD 2016*, Riva del Garda, Italy, 2016. URL https://api.semanticscholar.org/CorpusID:199525337.

Miroslav Malik, Sharath Adavanne, Konstantinos Drossos, Tuomas Virtanen, Dasa Ticha, and Roman Jarina. Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition. *ArXiv*, abs/1706.02292, 2017. doi: 10.48550/arXiv.1706.02292.

Konstantin Markov and Tomoko Matsui. *Speech and Music Emotion Recognition Using Gaussian Processes*, page 63–85. Springer Japan, 2015. ISBN 9784431553397. doi: 10.1007/978-4-431-55339-7_3.

M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler. Omras2 metadata project. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*. ISMIR, 2009. URL https://ismir2009.ismir.net/proceedings/LBD-18.pdf.

Merriam-Webster. Emotion. *Merriam-Webster*, 2023. ISSN 1754-0739. URL https://www.merriam-webster.com/dictionary/emotion.

L. B. Meyer. *Explaining Music: Essays and Explorations*. University of California Press, Berkeley, CA, USA, 1973. doi: /10.2307/jj.8501512.

Owen Craigie Meyers. *A Mood-Based Music Classification and Exploration System*. Msc, Massachusetts Institute of Technology, 2007. URL https://dspace.mit.edu/handle/1721.1/39337http://web.media.mit.edu/$\sim$meyers/meyers-ms.pdf.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997. URL https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf.

Oriol Nieto, Matthew C. McCallum, Matthew E. P. Davies, Andrew Robertson, Adam M. Stark, and Eran Egozy. The Harmonix Set: Beats, Downbeats, and Functional Segment Annotations of Western Popular Music. In *International Society for Music Information Retrieval Conference*, 2019. URL https://api.semanticscholar.org/CorpusID:208334350.

Oriol Nieto, Gautham J. Mysore, Cheng i Wang, Jordan B. L. Smith, Jan Schlüter, Thomas Grill, and Brian McFee. Audio-Based Music Structure Analysis: Current Trends, Open Challenges, and Applications. *Trans. Int. Soc. Music. Inf. Retr.*, 3:246–263, 2020. URL https://api.semanticscholar.org/CorpusID:230108966.

Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. Keras Tuner. https://github.com/keras-team/keras-tuner, 2019.

Saman Omer, Zakia Ghafoor, and Shavan Askar. Plant Disease Diagnosing Based on Deep Learning Techniques: A Survey and Research Challenges. *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, 11:38–47, 02 2023. doi: 10.14500/aro.11080.

Richard Orješek, Roman Jarina, and Michal Chmulik. End-to-end music emotion variation detection using iteratively reconstructed deep features. *Multimedia Tools and Applications*, 81, 02 2022. doi: 10.1007/s11042-021-11584-7.

H. Owen. *Music Theory Resource Book*. Oxford University Press, London, UK, 2000. URL https://discovered.ed.ac.uk/discovery/fulldisplay?vid=44UOE_INST:44UOE_VU2&search_scope=UoE&tab=Everything&docid=alma9916630683502466&lang=en&context=L&adaptor=Local%20Search%20Engine&query=sub,exact,%20Semiotics.

Renato Panda and Rui Pedro Paiva. Using Support Vector Machines for Automatic Mood Tracking in Audio Music. In *130th Audio Engineering Society Convention 2011 (AES)*, pages 579–586, London, UK, 2011. ISBN 9781617829253. URL https://hdl.handle.net/10316/95172.

Renato Panda, Bruno Rocha, and Rui Pedro Paiva. Music Emotion Recognition with Standard and Melodic Audio Features. *Applied Artificial Intelligence*, 29:313–334, 04 2015. doi: 10.1080/08839514.2015.1016389.

Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. Musical Texture and Expressivity Features for Music Emotion Recognition. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018. ISMIR. URL https://mir.dei.uc.pt/pdf/Conferences/MOODetector/ISMIR_2018_Panda.pdf.

Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. Novel Audio Features for Music Emotion Recognition. *IEEE Transactions on Affective Computing*, 11(4):614–626, 2020a. ISSN 1949-3045. doi: 10.1109/TAFFC.2018.2820691.

Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. Novel Audio Features for Music Emotion Recognition. *IEEE Transactions on Affective Computing*, 11(4):614–626, 2020b. doi: 10.1109/TAFFC.2018.2820691.

Alessia Pannese, Marc-André Rappaz, and Didier Grandjean. Metaphor and music emotion: Ancient views and future directions. *Consciousness and Cognition*, 44:61–71, 2016. ISSN 1053-8100. doi: https://doi.org/10.1016/j.concog.2016.06.015.

Sebastien Paquette, Isabelle Peretz, and Pascal Belin. The "Musical Emotional Bursts": A validated set of musical affect bursts to investigate auditory affective processing. *Frontiers in psychology*, 4:509, 08 2013. doi: 10.3389/fpsyg.2013.00509.

James Russell. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980. doi: 10.1037/h0077714.

Emery Schubert. Update of the Hevner Adjective Checklist. *Perceptual and Motor Skills*, 96(3_suppl):1117–1122, 2003. doi: 10.2466/pms.2003.96.3c.1117. PMID: 12929763.

Emery Schubert. Modeling Perceived Emotion With Continuous Musical Features. *Music Perception*, 21:561–, 06 2004. doi: 10.1525/mp.2004.21.4.561.

J. Smith, J. Burgoyne, I. Fujinaga, D. De Roure, and J. Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 555–560. ISMIR, 2011. URL `https://archives.ismir.net/ismir2011/paper/000099.pdf`.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 2nd edition, 2018. ISBN 9780262039246. URL `https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf`.

George Tzanetakis and Perry Cook. Marsyas: A framework for audio analysis. *Organised Sound*, 4(3):169–175, 12 2000. doi: 10.1017/S1355771800003071.

Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary Detection in Music Structure Analysis using Convolutional Neural Networks. In *International Society for Music Information Retrieval Conference*, 2014. doi: https://api.semanticscholar.org/CorpusID:9393056.

Sandrine Vieillard, Isabelle Peretz, Nathalie Gosselin, Stéphanie Khalfa, Lise Gagnon, and Bernard Bouchard. Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion - COGNITION EMOTION*, 22:720–752, 06 2008. doi: 10.1080/02699930701503567.

Ju-Chiang Wang, Yun-Ning Hung, and Jordan B. L. Smith. To catch a chorus, verse, intro, or anything else: Analyzing a song with structural functions, 2022.

Shuo-Yang Wang, Ju-Chiang Wang, Yi-Hsuan Yang, and Hsin-min Wang. Towards time-varying music auto-tagging based on CAL500 expansion. *Proceedings - IEEE International Conference on Multimedia and Expo*, 2014:1–6, 07 2014. doi: 10.1109/ICME.2014.6890290.

Jing Yang. A Novel Music Emotion Recognition Model Using Neural Network Technology. *Frontiers in Psychology*, 12, 2021a. ISSN 1664-1078. doi: 10.3389/fpsyg.2021.760060.

Jing Yang. A Novel Music Emotion Recognition Model Using Neural Network Technology. *Frontiers in Psychology*, 12, 2021b. ISSN 1664-1078. doi: 10.3389/fpsyg.2021.760060.

Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H. Chen. A Regression Approach to Music Emotion Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):448–457, 2008. doi: 10.1109/TASL.2007.911513.

Marcel Zentner, Didier Grandjean, and Klaus Scherer. Emotions Evoked by the Sound of Music: Characterization, Classification, and Measurement. *Emotion (Washington, D.C.)*, 8:494–521, 08 2008. doi: 10.1037/1528-3542.8.4.494.

Appendices

Appendix A

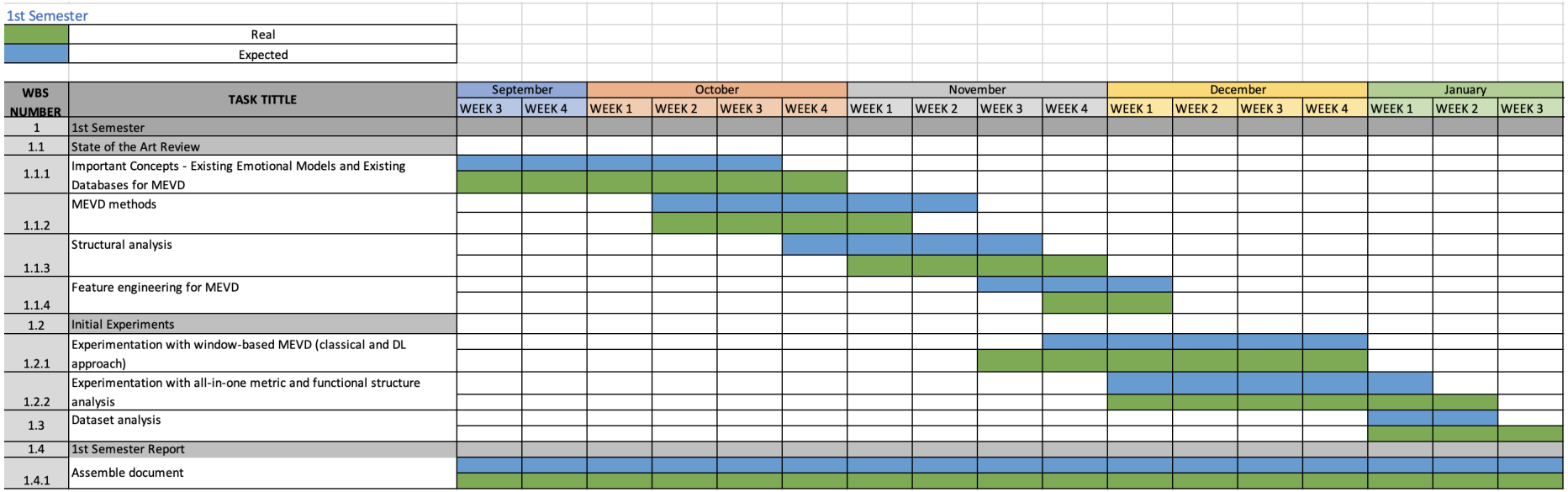


Figure A.1: Estimated and real effort for the first semester.

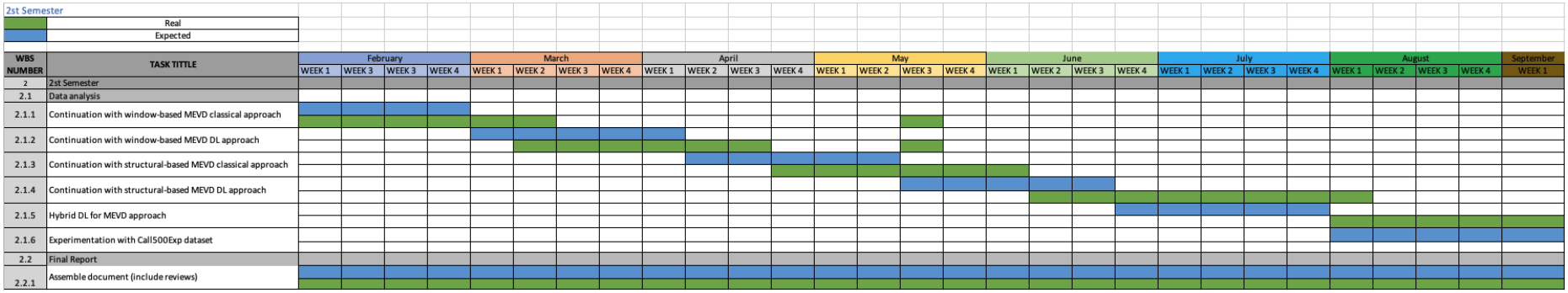


Figure A.2: Estimated and real effort for the second semester.